APPENDIX C

INVENTORY MODELS

The Class V Underground Injection Control Study

Appendix C
Inventory Models

September 30, 1999

U.S. Environmental Protection Agency
Office of Water
Office of Ground Water and Drinking Water
Implementation and Assistance Division

Table of Contents

Page

Table of Contents (continued)

1. General Approach to the Inventory Models

1.1 Overview of Class V Study

In 1997, the U.S. Environmental Protection Agency's (USEPA's) Office of Ground Water and Drinking Water (OGWDW) launched a national study of Class V Underground Injection Control (UIC) wells to determine whether additional regulations are needed to protect underground sources of drinking water. Class V wells are predominantly shallow injection wells that have a variety of uses, including disposal, aquifer recharge, and mineral recovery. The Agency has identified and collected information on 23 well subclasses, varying in complexity from shallow storm water drainage wells and large-capacity septic systems (LCSSs) to sophisticated geothermal reinjection wells.

USEPA conducted the Class V study in order to meet the requirements of the Safe Drinking Water Act (SDWA) and a modified consent decree with the Sierra Club Legal Defense Fund. USEPA will use the information collected in the study to help it determine whether additional regulations are necessary. USEPA will publish a final report on the study's findings in September 1999.

The study has two components: a general information collection for 23 subclasses of Class V wells and a model to estimate the number of LCSSs and storm water drainage wells. Development of the inventory model depended largely onsite visits to a selected number of census tracts. The site visitors interviewed state and local UIC officials and then counted the number of storm water drainage wells and LCSSs in the census tract. The information in this Appendix pertains to the inventory model.

1.2 General Information Collection Effort

USEPA initiated the general information collection effort by convening a workgroup of USEPA and state UIC representatives to help design the study. Workgroup members met during the spring and summer of 1997 to develop a methodology for collecting information.

The general data collection effort was geared at collecting existing Class V data from state agencies. We collected information on the number and location of existing wells, regulatory and permitting requirements, injectate quality studies, contamination incidents, best management practices, and studies linking Class V wells to changes in ground water quality.

Although we collected a significant amount of information in the state data collection effort, the effort showed that states were lacking data in many areas. For example, states generally have limited information on storm water and agricultural drainage wells (ADWs) and LCSSs. Although some data was gathered at the local level, this study generally targeted states, not local and county officials.

1.3 Need for Quantitative Models

Although the general data collection effort did include storm water wells and LCSSs, through the work group meetings, we knew that very little inventory information was available from the states. States generally believe that their inventories of these well types are inaccurate and would not provide a realistic national estimate. As a result, the workgroup determined that it would be necessary to construct inventory models to provide national estimates of the numbers of storm water drainage wells and LCSSs. The inventory models predict the number of wells nationally based on geologic, demographic, and other characteristics of the census tracts that make up the nation. These estimates can then be used as part of an overall assessment of the extent to which storm water drainage wells and LCSSs threaten public safety.

There is little theory -- and virtually no empirical research -- regarding the factors affecting the number and location of these wells. Therefore, the form of the models and their specifications depend in large part on the data collected through site visits to census tracts. It is in this sense that the modeling exercise is truly exploratory. It allows us to test its assumptions about the relationships between the geologic and demographic characteristics of a census tract and the number of wells in that tract.

1.4 Separate Treatment of Agricultural Drainage Wells

Existing data on ADWs also is inadequate to estimate the number of these wells nationwide. Most ADWs are located on private property making them generally inaccessible to site visitors. This problem of inaccessibility constrained our ability to collect data on this well type and to construct a model for it. This is discussed more fully in Section 4 below.

1.5 Outline of Analysis

The remainder of this document describes the analysis undertaken to estimate the number of LCSSs and storm water drainage wells. It first describes the steps taken to identify and select the sample of census tracts used to develop the inventory models. It then turns to a discussion of the approach used to collect the data, including a detailed discussion of the site visit methodology. Next, it presents the inventory models that were developed using the data from the sample, and presents the models' estimates of the number of LCSSs and storm water drainage wells. The details of the steps taken to develop the models are shown in technical attachments to this Appendix. Finally, the document turns to a discussion of our treatment of ADWs.

2. Selection of Census Tracts for Site Visits

The inventory models were designed to predict the number of LCSSs and storm water wells nationally based on the geologic, demographic, and other characteristics of specific census tracts. Using these variables, we picked 99 census tracts across the nation with varying characteristics. The tracts were selected to be representative of geologic and demographic characteristics nationwide. Site visitors visited these census tracts, obtained maps, talked to local officials, and drove the streets in an effort to enumerate the wells in each selected tract. This data was then used to model the number of LCSSs and storm water wells nationwide.

This approach was chosen to make the best use of the limited resources available to conduct this study. This decision was based on the data quality objectives process described in USEPA's *Guidance for the Data Quality Objectives Process* (USEPA, 1994). While a larger sample would increase the precision of the estimate of the number of wells, the costs of increasing the sample size outweighs the benefits. It would be prohibitively expensive to sample the 600-1,000 Census tracts necessary to produce precise estimates of the number of wells in the nation, both in terms of the financial costs of conducting the site visits, and the time required to complete the visits. Furthermore, the payoff to an increase in the precision of the estimate of the number of wells is relatively small. The workgroup determined that reducing sampling design and measurement errors by drawing a large sample was not necessary to meet the needs of the study. The goal of this analysis is not merely to produce an estimate of the number of storm water drainage wells and LCSSs in the country; rather, it is to assess the potential risks to public safety posed by these wells. This risk assessment is a complex endeavor, in which the number of wells is only one of several sources of uncertainty. For example, in the case of a single well, whether contaminants drained into the well work their way into the drinking water supply depends on many factors, including soil and bedrock characteristics and the type of contaminant. While contaminants may be found in the drinking water supply, there are myriad sources of contamination in addition to the well, so determining the threat posed by the well is difficult.

On the national level, the number of wells affect the overall risk, but even with a complete census of the number of wells in the nation, this uncertainty in the risk analysis would remain. Therefore, the goal of these models is to provide a reasonable estimate of the number of each type of well, which can inform the risk assessment. The risk assessment also can make use of the models' estimates of the variance around the estimated total to determine how sensitive the potential risks are to the estimate of the number of wells. The consensus of the workgroup was that a sample of 100 tracts would provide the information needed to inform the risk analysis, within the financial and time constraints faced by the study. (The sample consists of 99 rather than 100 tracts because the tracts were selected in 33 clusters of 3, as explained below.)

2.1 Steps for Selecting Census Tracts for Site Visits

The process used to select the 99 census tracts involved three steps. First, we identified tracts that are eligible to be included in the sample. In the second step, we stratified the tracts and selected 33 "target tracts." In the final step, we identified other tracts near the target tracts and formed clusters of three tracts each. This clustering approach reduced the travel required to collect data, while still allowing us to choose tracts that reflect a full range of characteristics. Throughout the process, tracts were chosen to reflect the full range of eligible tracts.

2.1.1 Identify Eligible Tracts

There are approximately 62,000 census tracts in the United States, including block numbering areas.[1] The first step in the development of the sample was to identify the tracts that were eligible to be included in the sample. Based on the consensus of the workgroup representatives, tracts that were predominately in federally-owned lands or in "urbanized areas," as defined by the U.S. Bureau of the Census, were excluded.[2] It was the workgroup's consensus that these areas would not contain LCSSs. While a small number of urbanized areas use storm water drainage wells, most do not. The consensus of the workgroup representatives was that a sample of tracts in urbanized as well as non-urbanized areas would not adequately represent the small number of urbanized tracts with storm water drainage wells. Also, some good data are available on the number of wells in some cities. Therefore, we decided to rely on the general data collection effort to enumerate the number of storm water drainage wells in urbanized areas, and to use the model to estimate only the number of these wells in non-urbanized areas. There are 24,818 Census tracts outside of federally-owned lands and urbanized areas.

Based on the consensus of the workgroup, we eliminated additional tracts from the list of eligible tracts that we thought were highly unlikely to contain either LCSSs or storm water drainage wells. Again, to the extent storm water drainage wells may be in some of these tracts, we will account for them through the general data collection effort. Tracts with the following characteristics were declared ineligible:

- *Tracts with high housing or population densities*. These tracts are largely near urbanized areas, and households in these areas are typically served by public sewer systems. Based on the workgroup consensus, we decided to exclude tracts in the top five percentiles of the housing and population distributions.

- *Tracts with very low population densities and/or very large areas*. These sparsely populated areas -- the bottom five percentiles of the population distribution (less than seven people per square mile) or the largest five percentiles by area (330 square miles or larger) -- are often in remote rural areas and are unlikely to have either subclass of well. We excluded these tracts. Again, this exclusion is based on the consensus of the workgroup.

- *Tracts with gypsum and salt bedrock*. These are tracts in which gypsum and salt deposits occur near the land surface and form karst-like features. This type of bedrock unit is not conducive to ground water movement, either to produce water or to accept water from an

[1] Some areas of the country are designated "block numbering areas" by the Census Bureau, rather than Census tracts. All are referred to as "census tracts" in this document.

[2] "Urbanized area" is defined by the U.S. Census Bureau as "a continuously built-up area with a population of 50,000 or more. It comprises one or more places – *central place(s)* – and the adjacent densely settled surrounding area – *urban fringe* – consisting of other places and nonplace territory."

injection well. When ground water is present, it is generally unpotable. Tracts classified as having this type of bedrock -- which are one percent of the tracts outside federal lands and urbanized areas -- were excluded.

- *Tracts that are largely crop land*. These are tracts in which a very large percentage of the area is farmland. These tracts may have other drainage wells, but they are not the focus of this site selection process. Based on expert opinion and workgroup comment, we excluded tracts in the top five percentiles of the distribution of the percentage of the area in crop land.

After eliminating these tracts, the number of eligible tracts was 18,578, or just over three quarters of the 24,181 tracts outside of federal lands and urbanized areas.

2.1.2 Select Target Tracts

After we identified eligible tracts, the next step was to select target tracts. We considered two options:

- *Random selection.* In this option, USEPA would select tracts at random, which would theoretically result in a set of tracts with a full range of demographic and geologic characteristics. The advantage of this option is that every tract would have an equal probability of selection. The disadvantage is that many of the site visits would likely occur in areas that contain few or no wells.

- *Stratified selection.* In this option, which the workgroup chose, we selected some tracts that are likely to contain wells and some tracts that are not, but we weighted the selection process toward the ones likely to contain wells. One advantage of this option is that it improves the cost effectiveness of the data collection. (Confirming, through many costly site visits, that wells do not exist, is expensive.) In addition, this option improves the statistical efficiency of the model.

There are two important notes about this site selection process. First, this approach produces a probability sample, although the probability of selection is not the same for each tract in the sample. We calculated the selection probabilities for each tract, and used these probabilities while developing the final model for the nation. In other words, the model does not assume that the areas we visited have the same characteristics, on average, as the areas we did not visit. Second, this methodology allowed for mid-course corrections in case our assumptions proved to be highly inaccurate.

The following describes the steps we used to stratify the census tracts into areas that are more likely to contain wells and areas that are less likely to contain wells. The first two variables are geologic characteristics that influence the probability that drainage wells are used -- the presence of the most susceptible bedrock conditions and/or the presence of extensive sand and gravel deposits. The final two variables deal with demographic characteristics that influence the number of wells.

1. *Distinguish between tracts with no susceptible bedrock from tracts with susceptible bedrock (as conditioned below) and presence of bedrock that is within five feet of the surface.*[3] Tracts that have the co-occurrence of susceptible bedrock and susceptible or other bedrock within five feet of the surface are distinguished from all other tracts. The susceptible bedrock category in the database is subdivided into seven classes.[4] The first four represent karst features for the United States; the last three are areas of "fractured" rock settings:

   A. Areas where normal fractured carbonate-rock derived karst topography, specifically flat lying limestones and dolostones, is at the land surface.

   B. Areas where normal fractured carbonate-rock derived karst topography, specifically moderately deformed carbonate rocks, is at the land surface.

   C. Areas where carbonate-rock karst is buried by up to 200 feet of noncarbonate material (e.g., glacial deposits in the upper Midwest).

   D. Areas where karst is formed in salt or gypsum beds.

   E. Deformed crystalline and carbonate rocks (marbles) affected by fracturing and solution ("karstification" in the carbonate portions). This unit is particularly important in portions of the Appalachian region as well as some of the western mountain areas.

   F. The fractured, porous volcanic rocks of the western United States, particularly the northwestern states (e.g., the Snake River aquifer and Columbia Plateau region).

   G. Large areas with known jointing and collapse features in sedimentary deposits (e.g., silt, sands, and gravels), particularly in the southern coastal Plain and High Plains region of Texas.

   For the purpose of selecting tracts for site visits, only the classes for traditional karst (A-C) and the areas of volcanic/lava with fissures, tubes, and tunnels (F) were included in this variable. As noted above, the areas of gypsum bedrock karst (D) were excluded because such rock units are not typically useable for injection (or withdrawal) of water. Other classes, such as sedimentary silts and sands with collapse features, were not used for targeting site selection but remain in the database for evaluation.

[3] USEPA will refer to this variable as "susceptible bedrock" throughout this document.

[4] Source: Digital/GIS version of the *Engineering Aspects of Karst.* 1986. U.S. Geological Survey.

2. *Distinguish between tracts with soils that contain a relatively large percentage of sand and gravel from other tracts.* Tracts in which less than ten percent of the area is sand and gravel were placed in one stratum. This includes 75 percent of the tracts. The remaining 25 percent (i.e., those with greater area of sand and gravel) were placed in a second stratum. Workgroup members generally agreed that Class V wells are less likely to occur in tracts in the first stratum (i.e., those in which less than ten percent of the area is sand and gravel) than in tracts in the second stratum (i.e., those with greater area of sand and gravel).

   We stratified along this variable so that we could examine areas with greater percentages of sand and gravel soils. The ten percent cutoff was chosen for two reasons. First, previous experience suggests that areas with greater than ten percent sand and gravel soils are more likely to be associated with sand and gravel deposits of significant thickness, and in turn, more likely to contain wells. Second, ten percent appears to reflect a natural cutoff, given the distribution of census tracts along this characteristic. The distribution of tracts by the percentage of the soil that is sand and gravel is highly skewed, with a long right tail. (See Figure 12.) This cutoff places the long tail in the second stratum.

3. *Stratify the data by the percentage of housing units in an area on a public sewer*. Census data on public sewers, which cover sanitary sewers and not storm sewers,[5] also were used to stratify the data. Areas in which a large share of the housing units -- 90 percent and above -- are on public sewer systems are less likely than other tracts to contain either LCSSs or storm water drainage wells. USEPA stratified the data along this characteristic to limit the number of tracts USEPA drew from areas with 90 percent or more of the area on public sewer systems (i.e., to enhance the number of tracts likely to have septic systems and storm water wells).

4. *Distinguish tracts with housing patterns representative of the country as a whole from less typical tracts.* Three attributes characterize these tracts as atypical:

   A. A large percentage of the housing structures in the tract contain five or more housing units; or

[5] The absence of information on areas that are served by storm sewers (but not by sanitary sewers) is a limitation of Census data. The Census of Population (Question 16) asks respondents if their household is connected to a public sewer; if not, respondents are asked to specify if their household is connected to a septic tank, cesspool, or other means of wastewater disposal. This means that the Census identifies only areas served by sanitary sewers or combined sewers; it does not provide data on areas served exclusively by storm sewers. It also does not identify areas where businesses may be served by sewers but households are not. Finally, USEPA is relying on the most recent Census, which is almost ten years old, and may undercount areas that are recently sewered. All of these limitations are balanced by the fact that Census data were uniformly collected from every Census tract. There is no other comparable data source.

B. A large share of the housing units in the tract are mobile homes; or

C. A large share of the housing units in the tract are part-time residences.

We considered the tract atypical if it fell into the top five percentiles of any of these three characteristics.[6] We believe that areas with these characteristics are likely to contain more wells than typical tracts. There are approximately 3,500 such tracts, or 15 percent of the tracts outside of federal lands and urbanized areas.

We included these atypical tracts in the list of eligible tracts and explicitly allow for the structural break in the data between typical and atypical tracts. This approach provides us with information about the tracts that are most likely to contain wells while limiting the impact these tracts have on their estimate of the number of wells in more representative areas.

The combination of these four variables gave us 16 strata from which to select target tracts. We randomly selected tracts from each stratum to ensure they included the full range of census tracts. To weight the sample toward tracts that were likely to have wells, we selected a higher proportion of tracts from areas that are unsewered and have susceptible bedrock. We selected 40 tracts to serve as potential target tracts. In some cases, potential targets are very close to one another. In cases in which two targets were adjacent to each other, we randomly dropped one of the targets. Seven targets were dropped, leaving us with the desired 33 target tracts. Table 1 shows the number of target tracts selected from each stratum, as well as the total number of eligible tracts in each of the 16 strata.

2.1.3 Selecting The Tracts to Visit

The final step was to identify the remaining tracts to visit. Our approach to selection was to identify clusters of tracts, thereby reducing travel time for the staff who visited the tracts. Each target tract was the center of a cluster. Two additional tracts were selected from all of the tracts within a 100-mile radius of the target tract. This yielded 99 tracts in 33 clusters.

We used systematic selection to pick the remaining two tracts for each cluster, rather than pure random selection. There were two constraints that limited our ability to use pure random selection. First, we wished to avoid to the extent possible selecting more than one tract from the same county.[7]

[6] The top five percentiles include tracts with 22 percent or more of their structures containing five or more housing units. Tracts in which 36 percent or more of the housing units are mobile homes fall into the top five percentiles of that variable. Tracts in which 29 percent or more of the housing units are part-time residences are in the top five percentiles of that variable.

[7] County ordinances and practices may be factors that predict the use of injection wells. Selecting several tracts in the same county therefore could introduce bias into the model.

Second, we wanted ensure that the sample reflects the full range of characteristics in the eligible tracts; therefore, we weighted the sample by the number of observations sought for each stratum.

Table 1. Number of Target Tracts per Stratum
(Number of Eligible Tracts are in Parentheses)

		Less than 90% of Area is on Public Sewer System		90 to 100% of Area is on Public Sewer System		Total
		Less than 10% of Area is S&G[1]	10% or More of Area is S&G	Less than 10% of Area is S&G	10% or More of Area is S&G	
Typical Tracts	No Susceptible Bedrock[2]	3 (6,984)	3 (2,682)	1 (1,394)	0 (290)	7 (11,350)
	Susceptible Bedrock within 5 Feet of Surface	3 (3,648)	4 (780)	0 (363)	1 (76)	8 (4,847)
Atypical Tracts	No Susceptible Bedrock[2]	4 (699)	3 (628)	0 (290)	1 (84)	8 (1,701)
	Susceptible Bedrock within 5 Feet of Surface	4 (374)	5 (207)	1 (64)	0 (15)	10 (660)
Total		14 (11,705)	15 (4,297)	2 (2,111)	2 (465)	33 (18,578)

[1] Sand and gravel.
[2] Absence of susceptible bedrock within five feet of the surface in the tract.

From a statistical standpoint, an essential feature of any random sample is that the sample has a known probability of selection. This feature is retained by this procedure and the probability of selection can be estimated.

2.2 Summary of Demographic and Geologic Characteristics of the Sample

Table 2 shows the number of tracts selected from each stratum. Table 3 summarizes the number of tracts in the sample selected from each USEPA Region, by strata. Every stratum is not represented in each USEPA Region. For example, there are no typical tracts in the sample from USEPA Region 2 that are from areas with no susceptible bedrock within five feet of the surface and sand and gravel soil in less than ten percent of the area. The inventory model is intended to give a reasonable estimate of the number of wells in the nation as a whole, not for individual USEPA Regions.

Table 2. Number of Selected Tracts per Stratum

		Less than 90% of Area is on Public Sewer System		**90 to 100% of Area is on Public Sewer System**		
		Less than 10% of Area is S&G[1]	**10% or More of Area is S&G**	**Less than 10% of Area is S&G**	**10% or More of Area is S&G**	**Total**
Typical Tracts	No Susceptible Bedrock[2]	13	11	2	1	27
	Susceptible Bedrock within 5 Feet of Surface	9	9	3	1	22
Atypical Tracts	No Susceptible Bedrock[2]	11	14	2	1	28
	Susceptible Bedrock within 5 Feet of Surface	11	9	1	1	22
Total		44	43	8	4	99

[1] Sand and gravel.

[2] Absence of susceptible bedrock within five feet of the surface in the tract.

Table 3. Number of Target Tracts Per USEPA Region by Strata

			90% & Below on Sewer		90 to 100 % on Sewer System		
		USEPA Region	Less than 10% of Area is S&G[1]	10% or More of Area is S&G	Less than 10% of Area is S&G	10% or More of Area is S&G	Grand Total
Typical Tracts	No Susceptible Bedrock[2]	1	2	1	0	0	3
		2	0	0	0	0	0
		3	1	1	1	0	3
		4	1	1	0	0	2
		5	3	4	0	0	7
		6	2	0	0	0	2
		7	2	0	0	0	2
		8	0	3	1	0	4
		9	2	1	0	1	4
		10	0	0	0	0	0
	Susceptible Bedrock[2]	1	1	0	0	0	1
		2	0	1	0	0	1
		3	1	2	0	1	4
		4	3	3	0	0	6
		5	1	2	0	0	3
		6	0	0	2	0	2
		7	2	1	1	0	4
		8	1	0	0	0	1
		9	0	0	0	0	0
		10	0	0	0	0	0
Atypical Tracts	No Susceptible Bedrock[2]	1	0	2	0	0	2
		2	1	1	0	0	2
		3	1	1	1	0	3
		4	1	3	0	0	4
		5	0	4	0	0	4
		6	1	0	0	0	1
		7	0	0	0	0	0
		8	1	1	0	0	2
		9	4	2	1	1	8
		10	2	0	0	0	2
	Susceptible Bedrock[2]	1	2	0	0	0	2
		2	0	0	0	0	0
		3	3	1	0	1	5
		4	1	5	1	0	7
		5	0	1	0	0	1
		6	1	0	0	0	1
		7	1	0	0	0	1
		8	1	0	0	0	1
		9	0	0	0	0	0
		10	2	2	0	0	4
Typical and Atypical Tracts with and without Susceptible Bedrock[2]		1	5	3	0	0	8
		2	1	2	0	0	3
		3	6	5	2	2	15
		4	6	12	1	0	19
		5	4	11	0	0	15
		6	4	0	2	0	6
		7	5	1	1	0	7
		8	3	4	1	0	8
		9	6	3	1	2	12
		10	4	2	0	0	6
		Total	44	43	8	4	99

[1] Sand and gravel.

[2] Susceptible bedrock within five feet of the surface in the tract.

Attachment C to this Appendix shows the distributions of several tract characteristics and compares the sample tracts with all eligible tracts. Each figure contains two graphs. The first graph is the distribution for the 18,578 eligible tracts, and the second graph is the distribution for the 99 tracts in the sample. Figure 11, for example, shows the distribution of the percentage of area with susceptible bedrock for the 18,578 eligible tracts and the 99 tracts in the sample. These two distributions show that the sample was drawn from the entire spectrum of eligible tracts and was weighted slightly toward those areas with the highest percentages of susceptible bedrock.

These data provide us with the probability sample that we used to develop the models of the inventory of storm water drainage wells and LCSSs. The modeling effort accounts for the weighted selection process used to choose the sample. Weights are assigned to each tract based on its selection probability. Because of the complex process used to draw this sample, the calculation of the selection probabilities is not trivial. These calculations are shown in Attachment B to this Appendix.

In studies of this sort, measurement error usually is the largest problem. In this study, the direction of measurement error seems clear. It is unlikely that field personnel reported wells that do not exist, but it is likely that they overlooked some wells that exist but are not evident through visual inspection. In short, measurement error is likely to contribute to an under-estimation of the number of wells. USEPA took several steps to try limit the amount of measurement error. Several sources were used to identify wells in each site visit, including state and local officials, local records, and physical inspections. Because the number of storm water drainage wells found during the site visits was fewer than expected, USEPA took two additional steps to attempt to verify its results. First, it sent two independent teams to inspect one site. The two teams identified the same number of wells in that site. Second, it selected nine additional sites for visits that the general data collection effort indicated were more likely to contain storm water drainage wells. More wells were in fact found in these tracts than in typical tracts in the sample, using the same methodology.

2.3 Site Visit Methodology

We conducted site visits to the 99 census tracts selected above to count the number of storm water drainage wells and LCSSs in each tract. State, local and USEPA Regional officials were contacted in advance of the visits for information on these well types and were invited to participate in the census tract site visits.

The process of locating storm water wells and LCSSs included interviewing USEPA, state, county, and local officials, state Department of Transportation engineers, city engineers, Drainage District commissioners, Natural Resource Conservation Service representatives, county engineers and sanitarians, and local Department of Highways and Roads engineers within a tract. In some cases, former UIC officials were also contacted. Site visitors obtained information on the design characteristics of storm water wells, the location of any known wells, areas likely to contain multiple grates leading to a single storm water well, and sewered and unsewered areas within each tract. Whenever possible, we obtained copies of local regulations and verified the existence of septic systems with permits or verbal confirmation from local officials.

Although contact with various officials was made prior to the site visits, face-to-face meetings were usually necessary in order to obtain detailed information. Typically, the site visitor would meet with county sanitarians and engineers prior to conducting the street survey. During these meetings, areas with the highest probability of containing LCSSs and storm water drainage wells would be discussed and identified on a local map. Visitors also obtained general design requirements as well as copies of any permits or sanitary and storm sewer plans on file.

We used 7.5 USGS Quad maps and census tract outlines as well as road, county, soil, and sewer maps to conduct field inspections of all streets in each census tract. The field inspections verified the accuracy of local officials' information and found wells not identified by local officials. We marked all roads traveled on the corresponding Quad maps. They usually were not able to enter private property, but stopped at all buildings that could potentially use a LCSS. We interviewed the owner or tenant of the building and/or obtained copies of permits, whenever possible. Site visitors typically stopped roadside to investigate local drainage patterns and the possibility of storm water drainage wells. They took a picture of structures they found and obtained the coordinates of the wells using hand-held global positioning system (GPS) units. At the end of each site visit, they compiled all data, permits, local regulations, daily logs, notes, pictures, and correspondence in a consistent format.

2.4 Results of Census Tract Site Visits

The census tract site visits produced the data needed to model the number of storm water drainage wells and LCSSs nationwide. LCSSs were found in 88.9 percent of the tracts visited. Storm water drainage wells were located in 22.2 percent of the census tracts surveyed. (See Figures 2 and 3.)

Storm water wells were primarily found along streets, but were also common in parking lots and in private residential compounds. A few storm water wells were found in other areas, such as along bike paths or in recreational vehicle parks, as shown in Figure 1. Demographic and geologic characteristics were shown not to be the sole determinants of the prevalence of wells. For example, adjacent counties sharing similar geological and demographic characteristics often differed in their use of storm water wells. In one county, site visitors found that hundreds of storm water wells were used, while the adjacent county used few if any storm water wells. State and county officials generally said that the cultural, political, and historical practices of a particular area strongly influenced the number and presence of storm water wells.

Figure 1.
Location of Storm Water Wells

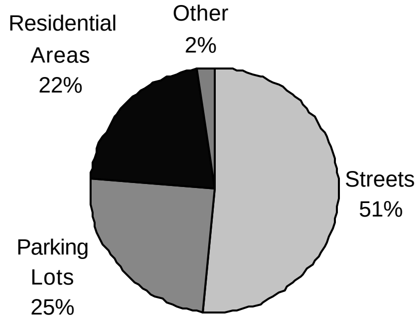


Figure 2. Location of LCSSs in Census Tract Visits

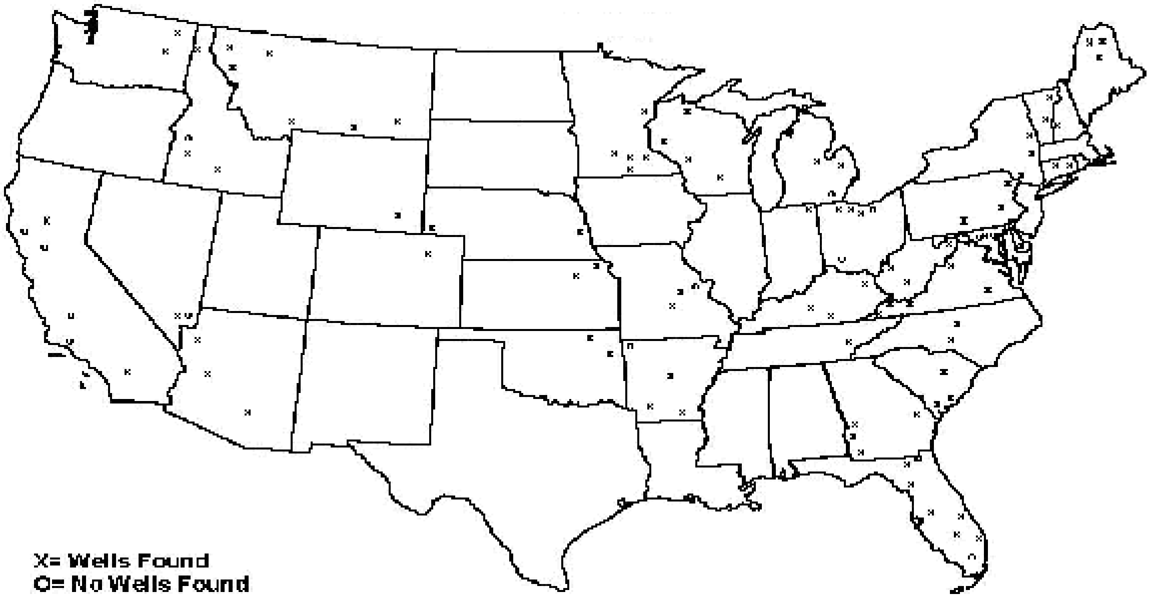
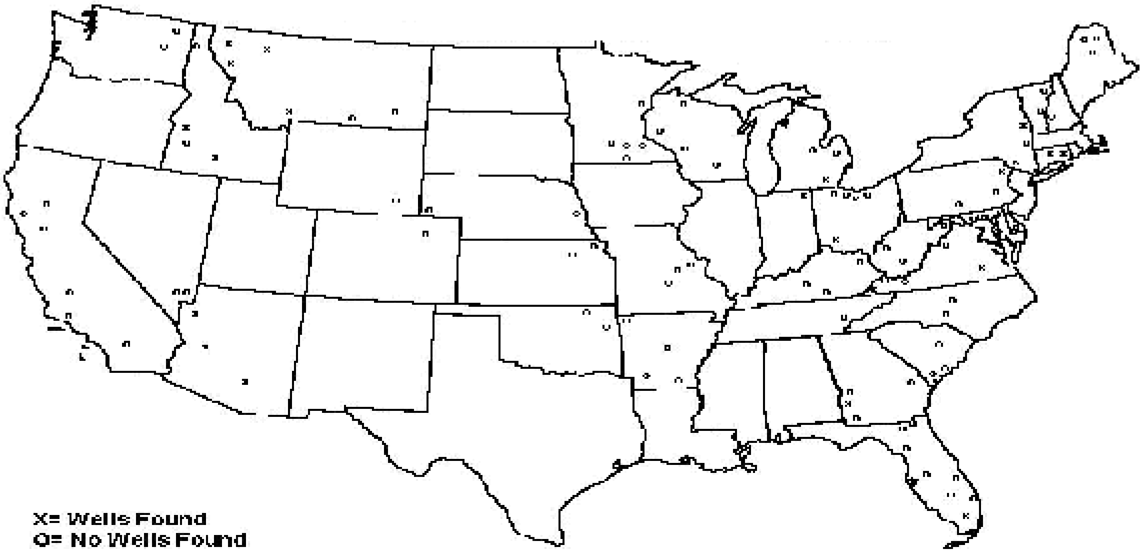


Figure 3. Location of Storm Water Drainage Wells in Census Tract Visits



LCSSs were found in sewered and unsewered areas and were used in a wide variety of circumstances. Geological variables did affect the existence and prevalence of septic systems. Septic systems were found in a wide variety of non-urbanized areas. As shown in Figure 4, the largest percentage of systems were located at churches, but there were also many found in commercial areas, restaurants, campgrounds, public buildings, motels, residential areas, industrial areas, schools, recreational areas, and a few in other areas such as farms and ranger stations.

Figure 4. Location of LCSSs

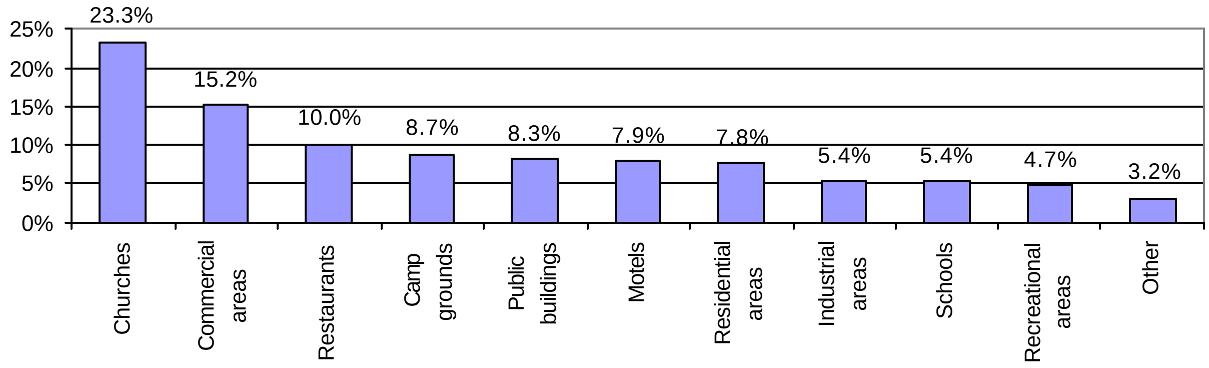


Table 4 summarizes the average number of storm water drainage wells found in the sample, by strata. Standard deviations are shown in parentheses. Due to the small sample size and the relatively few tracts that contain these wells, the differences are not statistically significant. Tracts in which 90 to 100 percent of the households are on public sanitary sewers have very few storm water drainage wells. Among tracts in which less than 90 percent of the households are on public sanitary sewers, the average number of storm water drainage wells is higher for tracts in which at least 10 percent of the area is sand and gravel soil. The difference is due largely to one tract, which had 210 wells. If this tract is excluded, the difference is much smaller. Atypical tracts also tend to have more wells if this one tract is excluded; this result is significant at the 10 percent level (but not at the 5 percent level).

Table 5 summarizes the results for LCSSs, by strata. Tracts in which 90 to 100 percent of the households are on public sanitary sewers are less likely to contain LCSSs. This is true in the aggregate, and within the other strata; the differences are statistically significant. Among tracts in which less than 90 percent of the households are on public sanitary sewers, there is little difference between tracts in which more than 10 percent of the soil is sand and gravel soil and other tracts, or between tracts with susceptible bedrock and tracts without susceptible bedrock. Atypical tracts tend to have more LCSSs, but the difference is not statistically significant at the 5 percent level.

Table 4. Average Number of Storm Water Drainage Wells in Sample by Strata
(Standard Deviations in Parentheses)

		Less than 90% of Area is on Public Sewer System		90 to 100% of Area is on Public Sewer System		
		Less than 10% of Area is S&G[1]	**10% or More of Area is S&G**	**Less than 10% of Area is S&G**	**10% or More of Area is S&G**	**Total**
Typical	No Susceptible Bedrock[2]	1.0 (3.6)	20.0 (63.0)	1.0 (1.4)	0.0 (NA)	8.7 (40.3)
	Susceptible Bedrock within 5 Feet of Surface	0.3 (0.7)	0.4 (0.9)	0.0 (0.0)	0.0 (NA)	0.3 (0.7)
Atypical	No Susceptible Bedrock[2]	4.4 (9.9)	2.7 (8.0)	0.5 (0.7)	0.0 (NA)	3.1 (8.3)
	Susceptible Bedrock within 5 Feet of Surface	7.5 (24.4)	0.9 (2.7)	0.0 (NA)	0.0 (NA)	4.1 (17.3)
Total		3.3 (13.2)	6.3 (32.2)	0.4 (0.7)	0.0 (0.0)	4.3 (22.9)

[1] Sand and gravel.
[2] Absence of susceptible bedrock within five feet of the surface in the tract.
NA - not applicable.

2.5 Data Limitations

While the data provide useful information about the number of storm water drainage wells and LCSSs, it is subject to certain limitations. Perhaps most importantly, the sample size is relatively small. Due to both financial and time constraints, the size of the sample selected for the analysis was limited to 100 Census tracts. While the stratification of the data along potential demographic and geologic variables ensured that the sample reflected the full range of these variables, the strata do not explain much of the variance in the number of either type of well. As a result, the standard errors of the estimated number of wells is relatively large, as will be seen in Section 3. The goal of the study is largely exploratory; very little is known about the number or location of either type of well or the factors that affect the decision to use these wells. USEPA was willing to accept some imprecision in order to develop initial estimates of the number of wells in the nation, and to gain understanding into the various factors that affect their use, while remaining within the constraints imposed on the study.

Table 5. Average Number of LCSSs in Sample by Strata
(Standard Deviations in Parentheses)

		Less than 90% of Area is on Public Sewer System		90 to 100% of Area is on Public Sewer System		Total
		Less than 10% of Area is S&G[1]	10% or More of Area is S&G	Less than 10% of Area is S&G	10% or More of Area is S&G	
Typical	No Susceptible Bedrock[2]	13.8 (10.1)	16.2 (18.9)	2.0 (2.8)	0.0 (NA)	13.4 (14.3)
	Susceptible Bedrock within 5 Feet of Surface	17.2 (15.8)	19.6 (15.0)	4.0 (6.9)	5.0 (NA)	15.8 (14.8)
Atypical	No Susceptible Bedrock[2]	33.1 (39.5)	21.6 (18.7)	2.0 (2.8)	2.0 (NA)	24.0 (28.9)
	Susceptible Bedrock within 5 Feet of Surface	15.3 (8.3)	23.1 (18.4)	13.0 (NA)	0.0 (NA)	17.7 (13.9)
Total		19.7 (22.7)	20.1 (17.5)	4.1 (5.5)	1.8 (2.4)	17.9 (19.8)

[1] Sand and gravel.
[2] Absence of susceptible bedrock within five feet of the surface in the tract.
NA - not applicable.

Census tract site visitors encountered various difficulties while conducting the site visits. Many wells were located on private or secured property, limiting site visitors' access to potential sites. At times, site visitors also had difficulty verifying information provided by local officials. Other limitations occurred because building owners and tenants were often not available, many parks and resorts operated seasonally, and certain buildings such as churches were only active during certain days or months.

In addition, site visitors found that state, county, and local officials could only provide limited information on storm water drainage wells and LCSSs. Many of their records were incomplete, missing, or out-of-date, which made it difficult to verify the wells' existence and configurations. For example, information was rarely available to help site visitors determine whether storm water drainage wells drained to the surface or ground water. Many officials were also reluctant to report the locations of known or suspected wells due to concerns that USEPA would later target them for enforcement action. Other problems encountered during the site visits included the inconsistent quality and availability of local maps, lack of daylight, and poor weather conditions which made inspection more difficult.

3. Models of Inventory of Large-Capacity Septic Systems and Storm Water Drainage Wells

The data collected through the site visits are used to develop models of the number of LCSSs and storm water drainage wells in the nation. This section describes these models and their estimates of the number of each subclass of well in the nation. Attachment A to this Appendix describes in detail how each model was developed. Based on these models, USEPA estimates the number of LCSSs to be approximately 300,000. The standard error of this estimate is approximately 20,000. We estimate the number of storm water drainage wells to be approximately 125,000. The standard error of this estimate is approximately 35,000.

3.1 Large-Capacity Septic Systems

The estimate of LCSSs was developed using a model-based approach. We used the sample of 99 tracts to estimate a model, which was then used to estimate the number of wells in each tract in the country. We estimated the model using Poisson regression. (See McCullagh and Nelder (1989) for a description of Poisson regression.) Poisson regression is an appropriate approach to use when modeling discrete counts, as we are in this case. The Poisson model estimates an *incidence rate* which it multiplies by an *exposure* to obtain the expected number of observed events. The exposure is the number of households on septic systems in a Census tract. The rate is a function of the characteristics of each Census tract, including the tract's housing density and the percentage of soil in the tract that is poorly drained. The model also distinguishes between tracts in which 90 to 100 percent of the households are on sanitary sewers and tracts in which fewer than 90 percent of the households are on sanitary sewers. It also allows the rate to vary by USEPA Region. Formally, the expectation of the number of LCSSs in an eligible Census tracts is given by equation 1:

(1)
$$E[LCS] = (Septic^{b_1+b_3 Sewered})*e^{\substack{b_0+b_2 Sewered+b_4 Density \\ +b_5 Drainage+\sum_{j=2}^{10} b_{3+j} EPA_j}}$$

Where: E[LCS] = the expectation of the number of LCSSs in a tract;

Septic = the number of households on sanitary septic systems in a tract;

Sewered = a dummy variable indicating 90 to 100 percent of tract is on public sewers;

Sewered*Ln(Septic) = an interaction term, which is the product of the Sewered dummy and the natural logarithm of the number of households on sanitary septic systems in a tract;

Density = households per square mile in a tract;

Drainage = the percentage of the tract with poor soil drainage;

EPA_j = nine dummy variables indicating USEPA Region 2 through USEPA Region 10 (USEPA Region 1 is the reference group); and

$\$_0$ through $\$_{14}$ are parameters estimated by a regression.

The term $Septic^{b_1+b_3 Sewered}$ is the exposure. It is multiplied by the rate, which is given by the term $e^{b_0+b_2 Sewered+b_4 Density+b_5 Drainage+\sum_{j=2}^{10} b_{3+j} EPA_j}$ to obtain the expected number of LCSSs per tract. Equation 1 can be restated as equation 2; the parameters of the model are estimated by running a Poisson regression on equation 2.

.

$$(2) \qquad E[LCS] = e^{\substack{b_0+b_1\ln(Septic)+b_2Sewered+b_3Sewered*\ln(Septic) \\ +b_4Density+b_5Drainage+\sum_{j=2}^{10} b_{3+j}EPA_j}}$$

While the parameters of the model are estimated with a Poisson regression, we do not know whether the true model is actually Poisson. The uncertainty about the underlying form of the model introduces another source of variability in our parameter estimates. The standard errors of our parameter estimates reflect this additional uncertainty. The coefficients are estimated assuming the Poisson model holds for both the mean and the variance; the standard errors are adjusted to allow for possible violations of these assumption about the variance. The standard errors are robust because they are consistent in large samples even if the data violate the assumptions used to produce estimates of the regression coefficients. (The standard errors are estimated based on the assumption that the observations are uncorrelated.) Put another way, if we are willing to assume that the data are distributed Poisson, we can eliminate a source of uncertainty and can reduce the standard errors of our estimates.

Table 6 displays the Poisson model's estimates of the parameters. Robust standard errors are shown in parentheses. The coefficient on the number of households on septic systems (the exposure variable) is 0.85, which means the exposure is not proportional to the number of households on septic systems. (The exposure would be proportional to the number of households on septic systems if the coefficient equaled 1.) The exposure increases 0.85 percent when the number of households on septic systems increase 1 percent, for tracts in which less than 90 percent of the households are on public sanitary sewers. This result is statistically significantly different from zero; it also is statistically significantly different from 1.

Tracts in which 90 to 100 percent of the households are on public sanitary sewers (which will be referred to as "sewered tracts") may behave differently than tracts in which less than 90 percent of the households are on public sanitary sewers (which will be referred to as "non-sewered tracts"). The intercept and the coefficient on the number of households on septic

Table 6. Poisson Model of LCSSs

	Parameter Estimates
Ln(Septic): Households on Septic Systems	0.849 (0.106)**
Sewered: 90-100% of Households on Sewers	1.320 (1.537)
Sewered*Ln(Septic)	-0.112 (0.327)
Density: Total Housing Units/Sq. Mile	-0.001 (0.001)
Drainage: Percentage of Area with Poorly Drained Soils	-0.010 (0.005)*
USEPA Region 2	1.018 (0.354)** †
USEPA Region 3	-0.007 (0.189)†
USEPA Region 4	0.546 (0.185)** †
USEPA Region 5	-0.424 (0.206)* †
USEPA Region 6	0.519 (0.339)†
USEPA Region 7	-0.413 (0.303)†
USEPA Region 8	0.960 (0.257)** †
USEPA Region 9	0.417 (0.202)* †
USEPA Region 10	-0.171 (0.315)†
Constant	-3.143 (0.811)**
Observations	99

Robust standard errors are in parentheses.
* Significant at 5% level; ** Significant at 1% level; † Jointly significant at 1 % level.

systems are allowed to be different for sewered tracts than for non-sewered tracts. This is accomplished by entering the dummy variable “Sewered” and the interaction term “Sewered*Ln(Septic).” The parameter estimate for the “Sewered” dummy variable is positive, but small, and the interaction term is negative. The positive sign of the dummy variable may seem counter-intuitive because it implies that the number of LCSSs per household on septic systems is higher in sewered tracts than non-sewered tracts, all else being equal. While this rate is higher, all else being equal, sewered tracts tend to have fewer LCSSs than non-sewered tracts for several reasons. First, sewered tracts tend to have far fewer households on septic systems than non-sewered tracts. Second, the negative sign on the interaction term’s coefficient means that the exposure is lower in sewered tracts than it is in non-sewered tracts, even if they contain the same number of households on septic systems. These two factors mean sewered tracts tend to have lower exposures than non-sewered tracts. Finally, the rate – the number of LCSSs per household on septic systems – may not be higher for sewered tracts than non-sewered tracts because all else is rarely equal. Housing density tends to be higher in sewered tracts than in non-sewered tracts, which, as we will see, tends to lower the number of LCSSs per household on septic systems. As a result of these three factors, the expected number of LCSSs in sewered tracts tends to be lower than in non-sewered tracts.

As housing density increases, the number of LCSSs decreases, as shown by the negative sign on the Density parameter in Table 6. The number of LCSSs also declines as the percentage of the tract with poorly drained soil increases. Neither of these results is statistically significant. The parameter estimates for each of the USEPA Region dummy variables are jointly significant, which means the average number of LCSSs per household on septic systems differs across the USEPA Regions.

The estimated number of LCSSs in a tract is given by equation 2 (or equation 1), using the values of the parameters shown in Table 6. The total number of LCSSs in the country is equal to the sum of the estimated number of systems across all tracts in the nation. The standard error of the estimate is equal to a linear combination of the parameters and their standard errors. (Attachment A to this Appendix shows the details of how we calculate the standard error.) A 95 percent confidence interval is equal to:

(3) $$CI = Total \pm 1.96*SE$$

Where: CI = 95% confidence interval;

Total = Poisson model's estimate of the total number of tracts in nation; and

SE = The standard error of the model.

The 95 percent confidence interval is defined such that if we repeatedly drew samples, estimated the model, and used the model to estimate the total number of LCSSs, this confidence interval would include the true answer 95 percent of the time. It assumes the variance in the estimate of the total number of wells is normally distributed.

As discussed in Section 2 above, the model assumed that relatively densely populated tracts surrounding urbanized areas would not contain wells. Areas with gypsum and salt bedrock are also excluded. The sample used to estimate the LCSSs model was drawn from the remaining 18,578 tracts in the 48 contiguous states. While site visits in Hawaii and Alaska could not be conducted, the model assumes that the parameters estimated for the rest of the country can be applied to those states as well. The total number of eligible tracts in the United States to which the model can be applied is 18,705. The model estimates that these tracts contain 289,385 wells, as shown in the first row of Table 7. The size of the 95 percent confidence interval is just under 100,000.

Table 7. Model-Based Estimate of the Number of LCSSs in the United States

	Estimated Number of LCSSs	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Eligible tracts in all states	289,385	21,165	247,902	330,868
Eligible tracts plus relatively densely populated tracts in all states [a]	353,361	25,143	304,081	402,641

[a] Includes eligible tracts, plus tracts in densely populated areas surrounding urbanized areas.

As discussed in Section 2, it was assumed that relatively densely populated tracts surrounding urban areas are highly unlikely to contain LCSSs. If this assumption is wrong, and these tracts contain wells, the model would underestimate the number of LCSSs in the nation. If it is assumed that the characteristics of these tracts affect the use of LCSSs in the same fashion as eligible tracts, the total number of wells increases by just over 60,000, to 353,361. This is shown in the second row of Table 7. The standard error also increases, as does the size of the 95 percent confidence interval. This estimate is meant to illustrate the potential impact of relaxing the assumption that these relatively densely populated tracts surrounding urban areas contain no wells. USEPA believes the assumption that these tracts contain few if any LCSSs remains valid, based on information collected during the general data collection effort and discussions with state and local authorities.

This estimate does not make use of information about the sampling design. As discussed in Attachment A to this Appendix, the regression could be weighted by the inverse of the selection probabilities. (See Attachment B to this Appendix for a description of how the selection probabilities are estimated.) This model-assisted approach would increase the estimate of the total number of LCSSs, and would increase the standard error of that estimate, as shown in Table 8. The standard error increases because of variability in the weights. This variability makes the use of weights less appealing because it increases the variance and likely does not affect bias as the strata do not appear to be associated with the number of LCSSs.

In addition to the sampling weights, the model could incorporate information about the stratification and clustering in the sample. We do not believe the data support the use of this information. The strata do not appear to be explain much of the variation in the data, so any adjustment would be small. Furthermore, the method used to estimate the standard errors that incorporate the information about the clustering can be highly variable when the number of clusters drawn from each strata is small (Kott, 1994). Furthermore, it is believed that the effect on the estimated standard errors of possible correlation among the observations is relatively small, see Attachment A to this Appendix.

Table 8. Model-Assisted Estimate of the Number of LCSSs in the United States

	Estimated Number of LCSSs	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Eligible tracts in all states	303,169	46,973	211,101	395,237
Eligible tracts plus relatively densely populated tracts in all states [a]	365,960	60,037	248,287	483,633

[a] Includes eligible tracts, plus tracts in densely populated areas surrounding urbanized areas.

3.2 Storm Water Drainage Wells

The estimate of the number of storm water drainage wells in the nation is itself the combination two estimates: a model estimate for wells in non-urbanized areas, and state estimates of the number of wells in urbanized areas. This approach is necessary because of the sampling strategy. Urbanized areas were excluded from the sample based on the assumption that very few storm water drainage wells would be found in urbanized areas. (See Section 2 above.) While a few cities make extensive use of these wells, they could not adequately be represented in our relatively small sample. Therefore, we decided to rely on state and other estimates gathered as part of the general data collection effort to account for the wells in urbanized areas, and to use the sample to build a model of the number of wells in non-urbanized areas. The estimate of the total number of wells in the country is the sum of these two estimates.

3.2.1 Two-Part Model of Storm Water Drainage Wells for Non-Urbanized Areas

The existence of storm water drainage wells in the sample is a relatively rare event. Of the 99 tracts in the sample, 22 contained storm drainage wells. Because a large share of the tracts contain zero wells, a linear model similar to the one used for LCSSs would be inappropriate. Therefore, a two-part model is used to estimate the number of wells in each tract. (See Duan et al., 1983 for a discussion of the two-part model.) The first part of the model estimates the probability that a given tract contains storm water drainage wells. The second part of the model estimates the average number of wells in tracts containing wells. The expected number of wells is then equal to the probability estimated in the first part of the model times the conditional mean estimated in the second part:

$$(4) \qquad E[SDW] = P(SDW > 0) * \overline{SDW}\Big|_{SDW>0}$$

Where: E[SDW] = the expectation of the number of storm water drainage wells in a given tract;

P(SDW>0) = the probability that tract contains storm water drainage wells; and

$\overline{SDW}\Big|_{SDW>0}$ is the average number of storm water drainage wells among tracts that contain wells.

Estimate the Probability that a Tract Contains Storm Water Drainage Wells

The probability that a tract contains a well is estimated using a probit regression. (See Aldrich and Nelson (1984) for a description of probit regression models.) The probability is estimated as a function of the density of housing built in the tract before 1970, the percentage of the area with poorly drained soils, and mean annual precipitation. The estimated probability is given by:

(5) $$P(SDW > 0) = \Phi\left(b_0 + b_1 Density_{70} + b_2 Drainage + b_3 MAP\right)$$

Where: M = the standard normal cumulative distribution function;

Density$_{70}$ is the number of housing units built before 1970, per square mile;

Drainage is the percentage of the tract with poor soil drainage; and

MAP is the mean annual precipitation in the tract, measured in inches.

Table 9 summarizes the results of the probit part of the model. The parameters are not shown because they are difficult to interpret; instead, the table shows the change in the probability of the existence of wells in a tract for a change in each explanatory variable (dF/dX), evaluated at the mean of the data.

The probability that a tract contains storm water drainage wells increases as the density of houses built before 1970 increases. If the number of houses per square mile built before 1970 in a tract increases by 10 from the observed mean, the probability of the tract containing storm water drainage wells increases by about one percent. The negative sign on the parameter for the percentage of the tract with poorly drained soil means the probability declines as the amount of poorly drained soil increases: a one percentage point increase in the amount of poorly drained soil in a tract reduces the probability of the tract containing storm drainage wells by about 0.8 percent. The implication is that storm water drainage wells may not be adequate to handle the run off in areas with poorly drained soils, so they are not used. This also may explain the sign on the mean annual precipitation parameter, which is negative as well. If precipitation increases by one inch, the probability that a tract contains storm water drainage wells decreases by about 0.7 percent.

Table 9. Probit Model of Probability of the Use of Storm Water Drainage Wells in Non-Urbanized Areas

	dF/dX
Density: Housing units built before 1970, per square mile	0.001 (0.001)*
Drainage: Percentage of tract with poorly drained soils	-0.008 (0.003)*
MAP: Mean annual precipitation	-0.007 (0.003)*
Observed probability	0.222
Predicted probability (at mean of data)	0.181
Observations	99

Standard errors in parentheses.
* Significant at 5% level.

Estimate the Average Number of Wells in Tracts that Contain Storm Water Drainage Wells

To estimate the number of wells in a tract, the predicted probability from the first part of the model is multiplied by the average number of wells in tracts containing wells. The site visits indicated that one factor that affects the number of wells in tracts containing wells is the existence of sand or gravel soil. As Table 10 shows, the average number of wells in tracts in which at least 10 percent of the soil is sand or gravel is considerably larger than tracts in which less than 10 percent of the soil is sand or gravel.

Table 10. Average Number of Storm Water Drainage Wells in Tracts Containing Wells

	Average Number of Storm Water Drainage Wells
Less than 10 percent of tract is sand or gravel soil	6.8
10 percent or more of tract is sand or gravel soil	18.2
All tracts	11.5

Note: Weighted average based on sample of 99 tracts. The number of tracts that contain storm water drainage wells is 22, 11 of which are in tracts in which less than 10 percent of the soil is sand or gravel.

The averages are weighted using the selection probabilities. (Weights are used because the simple mean does not include meaningful predictors, unlike the probit model of discussed above.) The predicted number of wells in a tract in which less than 10 percent of the soil is sand or gravel is equal to the product of the probability predicted by the probit model and 6.8. For tracts in which 10 percent or more of the soil is sand or gravel, the predicted number of wells is the product of the predicted probability from the unweighted probit model and 18.2. The estimated totals for the two-part model are shown in Table 11. As with LCSSs, the model distinguishes between eligible tracts and tracts in the more heavily populated areas surrounding urbanized centers. Unlike LCSSs, there is evidence that storm water drainage wells are used in some urbanized areas; therefore, it does not seem reasonable to assume that densely populated areas surrounding urbanized areas do not contain wells. If the model assumes these tracts behave like the eligible tracts included in the sample, the best estimate of the number of wells in non-urbanized areas is 63,998.

Table 11. Two-Part Model Estimate of the Number of Storm Water Drainage Wells in Non-Urbanized Areas in the United States

	Estimated Number of Storm Water Wells	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Eligible tracts in all states	44,246	24,670	9,630	103,008
Eligible tracts plus relatively densely populated tracts in all states [a]	63,998	35,278	13,409	145,174

[a] Includes eligible tracts, plus tracts in densely populated areas surrounding urbanized areas.

The error structure of the two-part model is relatively complex and it is difficult to estimate the standard error analytically. Therefore, the model uses a bootstrap method, described in Effron and Tibshirani (1993), to simulate the error structure and estimate the standard error. We drew 1,000 samples with replacement from the data, estimate the parameters of the two part model for each sample, and estimated the total number of wells nationally for each sample. These 1,000 estimates of the total have a mean and variance, which are used to estimate the standard error and the 95 percent confidence interval. The confidence interval is equal to the 2.5th and 97.5th percentiles of the 1,000 estimated totals. The standard error and confidence interval of the estimate of the total number of storm water drainage wells in non-urbanized areas is relatively large because of the relatively small size of our sample, and the relatively small number of tracts that contain wells.

The probit model used to produce the estimates in Table 11 does not incorporate the information about the sample design. If the probit model is weighted by the inverse of the selection probabilities, the estimated total number of wells in non-urbanized areas declines slightly, as shown in Table 12. As with the LCSS model, the model does not take the stratification and clustering into account when estimating standard errors. There is evidence that the possible correlation among the observations due to the clustering of the data has little impact on the estimated standard errors. (See Attachment A to this Appendix.)

Table 12. Weighted Two-Part Model Estimate of the Number of Storm Water Wells in Non-Urbanized Areas in the United States [a]

	Estimated Number of Storm Water Wells	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Eligible tracts in all states	40,455	23,952	8,006	95,491
Eligible tracts plus relatively densely populated tracts in all states [b]	59,105	34,107	11,345	131,600

[a] Probit model is weighted by the inverse of the selection probabilities.

[b] Includes eligible tracts, plus tracts in densely populated areas surrounding urbanized areas.

3.2.2 Storm Water Wells in Urbanized Areas

There are two sources of data on the number of storm water drainage wells in urbanized areas. The first are data sets from the states on the number and location of storm water drainage wells in those states. (Twenty-one states reported the use of wells in urban areas.) The location of municipalities in each data set was mapped to determine which ones fell into urbanized areas, as defined by the Census Bureau and this study. (See Section 2 above.) Approximately 35,000 wells were documented in urbanized areas. From the general data collection, states estimate an additional 27,000 wells in urbanized areas in the entire country. This is the second source of data used in this analysis. This is likely an underestimate, for several reasons. First, the states believe their estimates are lower than the actual number of wells that exist. Second, where a range was provided, we took the lower end of the range. And finally, it could not always be determined if the estimated number of wells was in urbanized or non-urbanized areas. Where this was the case, these estimates were not counted as part of the urbanized total. The estimate for the total number of wells in the country is equal to these estimates for urbanized areas plus the model's estimate for non-urbanized areas, which is approximately 125,000. This is shown in Table 13. The underestimate by the states would result in an underestimate of the total number of wells in the country.

Unlike the model estimate, there is no standard error associated with the data from the states. There is no standard error associated with document counts – they are fixed and known. USEPA cannot ascribe a standard error to the other estimates; thus, these estimates are treated as if they are fixed, known amounts. Therefore, all the variation in the estimate is due to the model, and none can be attributed to the data from the states. The 95 percent confidence interval around this estimate is 74,917 to 206,682. The confidence interval has a long right tail because of the small number of tracts with large numbers of storm drainage wells.

Of course, the other estimates of the number of wells are not fixed and known; therefore, the standard error of the estimate is greater than 35,278. If we assume the relative standard error of these estimates is similar to that of the two-part model (just over 50 percent), then the standard error of the estimated total would increase to just under 50,000. The 95 percent confidence interval also would increase; without further information about the distribution of the error inherent in the state estimates, it is not possible to determine the 95 percent confidence interval for the estimated total.

Table 13. Total Number of Storm Water Wells in Urbanized and Non-Urbanized Areas in the United States

	Estimated Number of Storm Water Wells	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Eligible tracts plus relatively densely populated tracts in all states [a]	63,998	35,278	13,409	145,174
Documented wells in urbanized areas	34,985	NA	NA	NA
Other urban wells	26,523	NA	NA	NA
Total	125,506	35,278	74,917	206,682

[a] Includes eligible tracts, plus tracts in densely populated areas surrounding urbanized areas.
NA = not applicable/unknown.

The states reported between 115,000 and 240,000 storm water drainage wells in the general data collection effort. Our estimate is toward the lower end of that range. The upper range of this estimate is not inconsistent with the upper range of the estimates from the states.

3.3 Conclusions

We estimate there are 289,000 LCSSs in the country. The 95 percent confidence interval is from approximately 250,000 to 330,000. It also estimates there are 125,000 storm water drainage wells, 60,000 of which are in a relatively small number of urbanized areas. The 95 percent confidence interval is from 75,000 to 207,000. This confidence interval reflects only the mean square error of the model of storm water drainage wells in non-urbanized areas. The number of wells in urbanized areas includes estimates by the states whose error cannot be quantified.

While the standard errors are relatively large (at least for the estimate of the number of storm water drainage wells), the estimates and confidence intervals are informative. It provides a general sense of the number of wells in the country. LCSSs are relatively common, but in relatively small numbers per census tract. Storm water drainage wells are used far less often, and are fewer in number. The size of the standard errors of these estimates reflects the relatively small size of the sample, and the wide range of factors that affect the use and number of these wells, many of which could not be measured in this study. The small number of tracts with relatively large numbers of storm water drainage wells also account for much of the imprecision of the model's estimate of the number of wells. If there are relatively few tracts in the county with large numbers of these wells, the national total would be towards the lower end of the range. If, on the other hand, there are a relatively large number of tracts in the nation with large numbers of storm water drainage wells, the number of wells in the nation could increase dramatically.

As discussed in section 2, both the estimate of the total number of wells and the estimate of the variance around that total may serve as inputs into an assessment of the risk posed by these wells. This risk assessment can determine whether the potential risk to public safety changes over the 95 percent confidence interval surrounding the estimated totals. To the extent the risk assessment changes dramatically over the 95 percent confidence intervals predicted by the models, further research on the number of wells in the nation may be warranted.

The models provide insight in addition to the estimate of the number of wells in the nation. The models indicate that the assumptions made when stratifying the eligible tracts are unsupported by the data. While there is some evidence that the amount of sand and gravel soil may affect the number of storm drainage wells, the evidence is weak. The existence of susceptible bedrock within five feet of the surface was not related to the number of either type of well; tracts with very high percentages of mobile homes, part-time residences, and structures with five or more housing units contained no more wells than more typical tracts when all other variables are held constant. Whether the households in a tract are largely on public sanitary sewers did affect the number of LCSSs; more generally, the number of households on public sanitary sewers was positively related to the number of LCSSs. The impact of the number of households on sewer systems on the number of storm water drainage wells was statistically insignificant. Other characteristics do affect the number of both type of wells, including soil drainage, precipitation, and housing density. While geologic characteristics like the prevalence of susceptible bedrock or sand and gravel soil may affect the feasibility of these wells, these other characteristics proved to be far more important. Other characteristics of tracts that are difficult to measure, including historical, cultural, and political factors, also are important. Their importance is reflected in the differences observed across regions of the country which remain after taking other characteristics into account, and in the amount of the variance in the number of wells that remains unexplained by the models.

One of our major concerns about the model's estimates is the extent to which it undercounted the number of wells in the census tracts it visited to develop the sample. As stated, USEPA took several steps to mitigate the effects of this measurement error. To consider how significant an impact the measurement error has on USEPA's estimates, USEPA compared the storm water drainage well model's results to the number of documented wells in ten states. USEPA compared the estimates for storm water drainage wells because it believes the potential undercount is larger for these wells than LCSSs, which were easier to identify. Arizona, Hawaii, Idaho, Indiana, Maryland, Michigan, Minnesota, Oregon, Utah, and Washington each provided data on the number and location of documented storm water drainage wells in their states.

One note of caution about this comparison. The storm water drainage well model is designed to estimate the number of wells in non-urbanized areas in the nation as a whole. These 10 states are not necessarily representative of the country as a whole, because they were not selected randomly.

They are relatively diverse geographically, geologically, and demographically, so the comparison is warranted.

The model estimates that approximately 21,000 wells are located in non-urbanized areas in these ten states. The states documented approximately 51,000 wells in both urban and non-urban areas; approximately 24,000 of these wells are in non-urban areas. This indicates that the model's estimate of the number of wells is reasonably accurate.

While the general data collection effort and the site visits provided us with unprecedented amounts of information about both storm water drainage wells and LCSSs, much remains unknown. Many of the assumptions made in the development of the sample were unsupported by the data, which reflects the general lack of understanding about the decisions regarding the use of the wells. The states' ability to track the use of these wells is inconsistent and often limited. The uncertainty surrounding the use of storm drainage wells and LCSSs is reflected in both the large range in the estimates provided by the states, and the large standard error of the models' estimates. Additional research can improve the precision of the estimates by developing additional data and by using the lessons learned in this study about the factors that affect the use of wells to guide the analysis.

4. Agricultural Drainage Wells

State inventories of ADWs are generally incomplete and underestimate the number of wells that may exist. Traditionally, states have been unable to keep accurate records of these wells because many of them exist on private property.

Initially, we considered using the site visit methodology described in this document as a basis for a model to estimate the number of ADWs. Site visits to locate ADWs would require a substantial time commitment and present potential difficulties because of the need to access private property.

Instead, we proposed to target geographical areas for review based on an analysis of existing information (e.g., from literature reviews, work group input, expert consultants). USEPA contacted state and local officials by telephone to describe the study, learn about current and historical agricultural practices in the area, determine the types of ADWs used in the area (e.g., flood irrigation return flow or wetlands drainage), learn how these practices are regulated, and identify the number and location of existing ADWs.

Through the general data collection effort, some location data was provided by states and incorporated into the state summaries of Class V wells. For example, Iowa provided detailed documentation of the wells including studies and inventories. However, many other states had very little data on this well type, although most suspect that the wells exist in the state. Other states were reluctant to report the locations of known ADWs due to concerns that USEPA would target them for enforcement action.

States that identified significant numbers of these ADWs in the general data collection effort included New York, Illinois, Minnesota, Ohio, Texas, Iowa, California, and Idaho. States that reported fewer than a hundred documented ADWs included Delaware, Indiana, Michigan, Wisconsin, Oklahoma, Oregon, Washington, and USEPA Region 10 Tribes. These states usually could not give an accurate estimate of this well type and predicted that more wells may exist in some areas. For example, Puerto Rico, Pennsylvania, Florida, and Kentucky could not provide any information on ADWs, but the states believe they exist. The data collection efforts on ADWs were constrained by the limited data available from states. We also made numerous phone calls to county and local officials to collect additional data, but found the same difficulties. Most ADWs are located on private property, making it extremely difficult for states and counties to locate them. Additional information on ADWs can be found in the corresponding volume of this Class V Study.

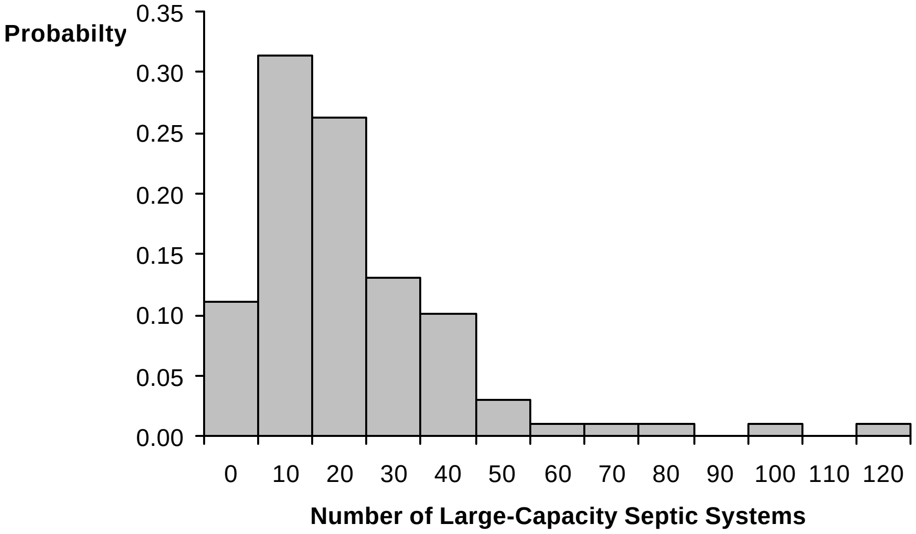
ATTACHMENT A
LARGE-CAPACITY SEPTIC SYSTEMS AND STORM WATER DRAINAGE WELLS INVENTORY MODELS

Section 3 of this Appendix presents the inventory models of LCSSs and storm water drainage wells. This attachment provides a technical explanation of how we developed each model. In doing so, it describes some of the assumptions behind each model, and explores the implications of relaxing those assumptions. It begins with a discussion of the LCSS model, and then turns to storm water drainage wells.

A.1 Large-Capacity Septic Systems

The average number of LCSSs in tracts in our sample is 18. Eleven tracts contain no LCSSs, and one tract has 119 systems. Figure 5 shows the probability distribution of LCSSs in the sample.

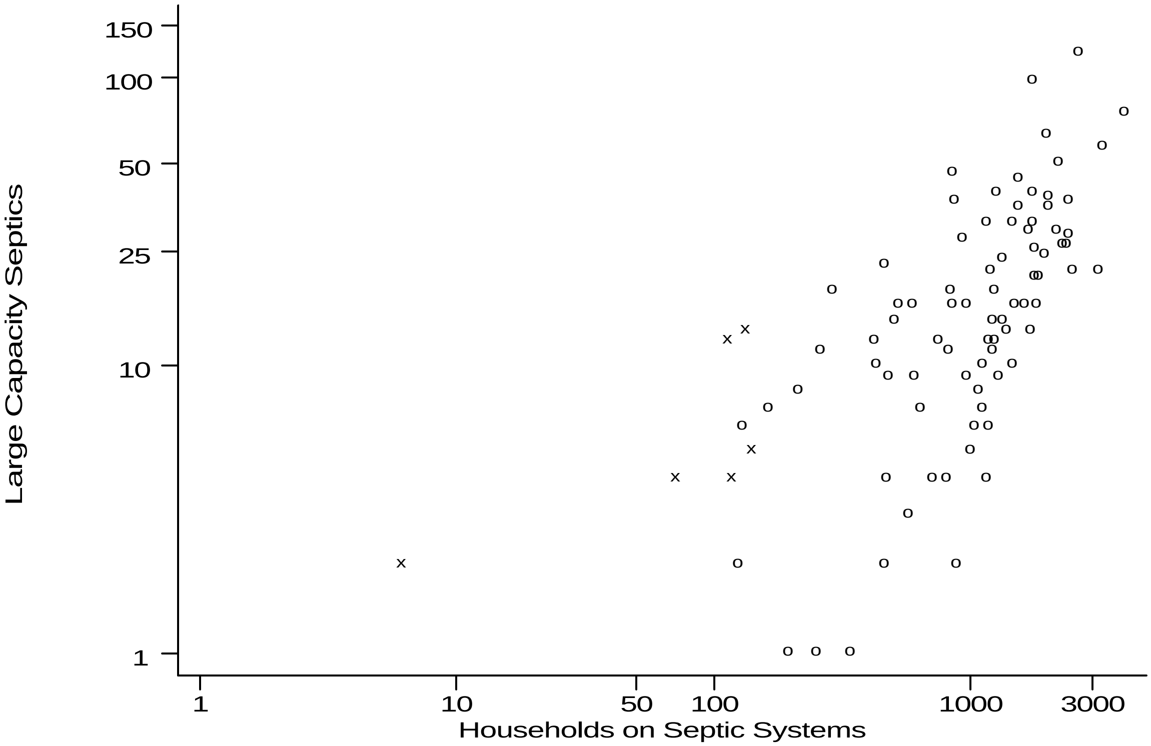
Figure 5. Distribution of LCSSs in Sample of 99 Non-Urbanized Census Tracts



We estimate a model in which the number of LCSSs in a given census tract is a linear function of the demographic and geologic characteristics of that tract. The number of LCSSs is a discrete distribution – the number can only be an integer; therefore, we modeled the relationship between the number of wells in a tract and the tracts characteristics with a Poisson regression (McCullagh and Nelder, 1989). The Poisson model assumes the errors are distributed Poisson. The Poisson distribution is a discrete distribution, in which the mean is equal to the variance.

A second characteristic of the Poisson model made it a good candidate for the LCSSs inventory model. The Poisson model predicts the occurrence of an event as a rate: it estimates the number of events per an exposure. The data in the sample indicate that a strong relationship exists between the number of households on septic systems and the number of LCSSs in a tract. For each household on a septic system in a tract (the exposure), there will be a certain number of LCSSs (the event). The intuition behind this approach is: (1) some of the households on septic systems will be on LCSSs, and (2) the more households there are in a tract on septic systems, the more likely it is that other buildings – schools, community buildings, religious establishments – will be on septic systems as well. Because these buildings are public buildings, they are likely to require LCSSs. Figure 6 plots the log of the number of LCSSs against the log of the number of households on septic systems. The number of LCSSs increases as the number of households on septic systems increases, as expected.

Figure 6. Count of LCSSs and Households on Septic Systems in Sample of 99 Census Tracts



The Poisson model estimates the number of LCSSs per household on septic systems (the rate) as a function of the characteristics of the tract. Poisson regressions often constrain the coefficient on the exposure variable to be equal to 1. Rather than impose this constraint, the model estimates the coefficient. This is accomplished by entering the natural logarithm of the number of households into the regression. The general form of the Poisson model is given by equation 6:

(6) $$E\left[LCS_i\right] = e^{\;0 + b_1 \ln(\,Septic_i\,) + \sum_{j=2}^{k} b_j X_{ij}}$$

Where: E[LCS$_i$] = the expectation of the number of LCSSs in tract i.

Septic$_i$ = the number of households on septic systems in tract i. This is the exposure, or the number by which the incidence rate will be multiplied to get the count for tract i. The natural logarithm of Septic is used by the model.

X_i = characteristics of tract i. The exponentiation of the sum of the intercept $\$_0$, and $\sum_{j=2}^{k} b_j X_{ij}$ equals the incidence rate for tract i.

$\$_1$ through $\$_k$ are parameters estimated by a regression.

The expectation of the number of LCSSs in tract i is equal to the exposure times the rate. This can be seen by re-arranging equation 6:

$$\text{(6a)} \qquad E\left[LCS_i\right] = \left(Septic_i^{\,b_1}\right) * \left(e^{b_0+\sum_{j=2}^{k} b_j X_{ij}}\right)$$

Tracts in which 90 to 100 percent of the households are on public sanitary sewers, which are marked with an "x" in Figure 6, tend to have fewer households on septic systems. The basic model assumes the incidence rate is different for these tracts. Using the dummy variable "Sewered" to identify these tracts, the basic model is given by equation 7:

$$\text{(7)} \qquad E[\,LCS\,]_i = e^{b_0+b_1\ln(Septic_i)+b_2 Sewered}$$

The results of this basic model are shown in column 1 of Table 14. The coefficient on ln(Septics) is 0.984, which is not statistically different from 1.0; in other words, the number of large scale septic systems is roughly proportional to the number of households on septic systems in this basic model. The coefficient on the Sewered dummy variable is positive, which indicates the incidence rate is higher for tracts in which 90 to 100 percent of the households are on public sanitary systems. The dummy variable is statistically significant at the one percent level. This somewhat counterintuitive result is due to the simple nature of this model. As we will see, the number of LCSSs is in fact lower in tracts in which 90 to 100 percent of the households are on public sanitary sewers.

An interaction term is introduced to allow the coefficients on both the exposure variable Ln(Septics) and the rate to be different for tracts in which 90 to 100 percent of the households are on public sanitary sewers. The results are shown in column 2 of the Table 14. The dummy variable and the interaction term are jointly significant at the one percent level, indicating that tracts in which households are overwhelmingly on public sanitary sewers do behave differently in our sample from other tracts. While the positive sign on the sewer dummy variable implies tracts in which 90 to 100 percent of the households are on public sanitary sewers contain more LCSSs per household than other tracts, all else remaining equal, this generally is not the case. The negative sign on the interaction term means the exposure is smaller for tracts in which 90 to 100 percent of the households are on public sewers than it is for tracts with fewer households on public sewers. Also, most tracts in which 90 to 100 percent of the households are on public sanitary sewers have relatively few households on septic systems. The net result is that the predicted number of LCSSs in tracts in which 90 to 100 percent of the households are on public sanitary sewers tends to be lower than in tracts in which less than 90 percent of the households are on public sanitary sewers. As we will see, other factors tend to reduce the number of LCSSs per household on septic systems for tracts in which 90 to 100 percent of the households are on public sanitary sewers.

Table 14. LCSS Inventory Models

	(1)	(2)	(3)	(4)	(5)	(6)
Ln(Septics)	0.984 (0.041)**	0.991 (0.042)**	0.991 (0.043)**	0.980 (0.042)**	0.980 (0.043)**	0.849 (0.044)**
90-100% Sewered	0.969 (0.192)**	2.093 (1.027)* †	2.312 (1.010)* †	2.045 (1.036)* †	2.261 (1.019)* †	1.320 (1.131)†
Sewered*Ln(Septic)		-0.236 (0.213)†	-0.252 (0.208)†	-0.229 (0.215)†	-0.242 (0.210)†	-0.112 (0.236)†
Total Housing Units/Sq. Mile			-0.001 (0.000)*		-0.001 (0.000)*	-0.001 (0.000)**
% Area Poorly Drained				-0.003 (0.002)*	-0.003 (0.002)*	-0.010 (0.002)**
USEPA Region 2						1.018 (0.125)** ‡
USEPA Region 3						-0.007 (0.114)‡
USEPA Region 4						0.546 (0.103)** ‡
USEPA Region 5						-0.424 (0.127)** ‡
USEPA Region 6						0.519 (0.136)** ‡
USEPA Region 7						-0.413 (0.197)* ‡
USEPA Region 8						0.960 (0.112)** ‡
USEPA Region 9						0.417 (0.125)** ‡
USEPA Region 10						-0.171 (0.164)‡
Constant	-3.992 (0.303)**	-4.044 (0.307)**	-4.011 (0.313)**	-3.926 (0.312)**	-3.889 (0.318)**	-3.143 (0.337)**
Observations	99	99	99	99	99	99

Standard errors are in parentheses.
* significant at 5% level; ** significant at 1% level; † these variables are jointly significant at 1% level; ‡ these variables are jointly significant at 1% level.

We next consider the impact of housing density and soil drainage on the model. We add the total housing density – the number of housing units per square mile – to the model. We expect the number of LCSSs per household to decline with housing density, as we believe more densely housed areas are more likely to be on public sanitary sewers. The probability distribution of housing density is shown in Figure 7. The distribution is skewed to the right, with two tracts containing over 600 housing units per square mile. The majority of tracts have fewer than 100 households per square mile. We also expect the number of LCSSs per household to decline as the percentage of the tract with poorly drained soil increases as the efficacy of the septic system declines with poor soil drainage. The distribution of the percentage of area with poorly drained soils is shown in Figure 8. It also is skewed to the right, as the percentage of the area with poorly drained soils is less than 20 percent in most tracts.

Figure 7. Distribution of Housing Density in Sample of 99 Census Tracts

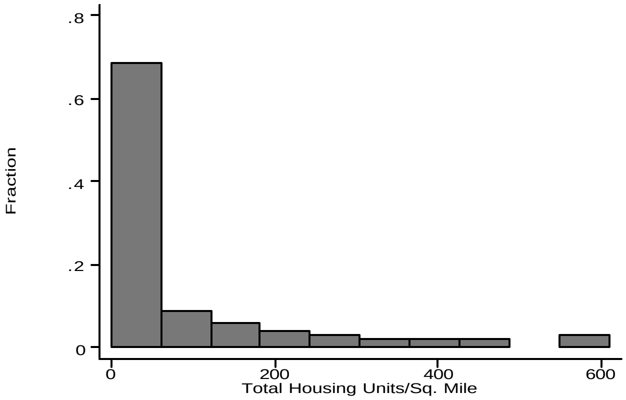
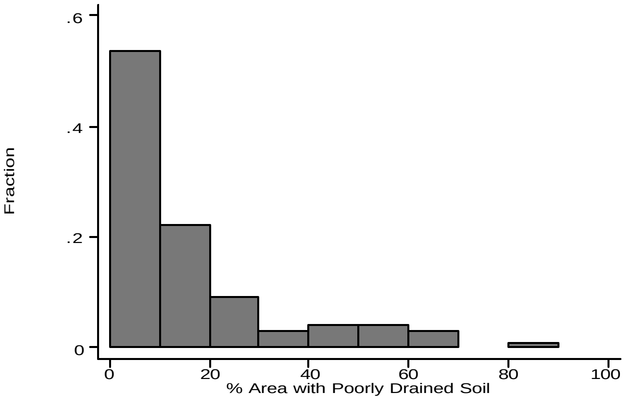


Figure 8. Distribution of Percentage of Area with Poorly Drained Soil in Sample of 99 Census Tracts



The results of the addition of housing density to the model are in column 3 of Table 14. The parameter's coefficient is negative and statistically significant. It implies that the number of LCSSs per household declines as housing density increases, as expected. In column 4 we add the impact of soil drainage on the use of LCSSs, ignoring the impact of housing density. As the percentage of the area with poorly drained soil increases, the number of LCSSs per household declines, as expected. Column 5 shows the model with both the housing density and soil drainage characteristics.

Finally, we test to see if the relationship between the number of LCSSs per household is the same across each of the ten USEPA Regions. The impact may vary due to regional differences, as well as other characteristics we cannot measure. Nine dummy variables are entered, for USEPA Region 2 through Region 10; USEPA Region 1 is the reference group. The coefficient on each dummy shows the difference between that USEPA Region and USEPA Region 1, holding all other variables constant. Using an F test of whether the Regional dummy variables are jointly equal to 0, we reject the hypothesis that they are equal to zero. This indicates the average number of LCSSs per household in each USEPA Region is different, holding all other characteristics constant. This version of the model is shown in the last column of Table 14; we will refer to it as the full model. (The Regional dummy variables allow us to incorporate the effect of regional differences in our model. It should be noted that this specification of the model does not necessarily allow us to predict the number of wells in each USEPA Region. Representative samples were not drawn from each USEPA Region – the sample is representative of the nation as a whole – and potentially important regional factors may be excluded from the model. While these factors may average out in the national sample, they could yield biased estimates for any single region.)

We also considered other explanatory variables. These included the percentage of the area with susceptible bedrock, the percentage of the area with sand or gravel soil, the percentage of the housing units that are mobile homes, the percentage of structures that contain five or more housing units, and the percentage of housing units that are part-time residences. None of these variables were either individually or jointly significant. In fact, of the variables used to stratify the eligible tracts in the country, the only one that results in a statistically significant effect is the sewered variable, which distinguishes tracts in which 90 to 100 percent of the households are on public sanitary sewers from those in which less than 90 percent of the households are on public sanitary sewers. Because the distributions of the housing density and soil drainage variables are skewed, we also considered transforming these variables. The transformations had little impact on the model.

In the full model, the coefficient on the exposure variable Ln(Septic) drops to 0.85, which is statistically different from 1.0 at the 1 percent level. Thus, the impact of the number of households is less than proportional for tracts in which less than 90 percent of the households are on public sanitary sewers. For tracts in which 90 to 100 percent of the households are on public sanitary sewers, the parameter for Ln(Septics) is even lower, although the difference is not statistically significant.

We use the full model to predict the number of LCSSs in each tract. The estimated number of LCSSs is given by equation 8:

$$(8) \qquad E[LCS] = e^{\substack{b_0 + b_1 \ln(Septic) + b_2 Sewered + b_3 Sewered * \ln(Septic) \\ + b_4 Density + b_5 Drainage + \sum_{j=2}^{10} b_{3+j} EPA_j}}$$

A.1.1 <u>Exploring the Implications of the Model's Assumptions</u>

An assumption underlying this model is that the data are distributed Poisson. This assumption can be formally tested. The test statistic is a P^2 with 84 degrees of freedom. The P^2 is 450; with a critical value of 108, we reject the hypothesis that the data are distributed Poisson. While the mean is equal to the variance in the Poisson distribution, the variance exceeds the mean in our data.

The full model was based on the assumption that the number of LCSSs in the population as a whole was generated by a Poisson model. Rejection of the hypothesis that the underlying data are distributed Poisson is an indication that a pure model-based approach is inappropriate, and that we must relax our assumptions regarding the underlying distribution of the data. Unfortunately, the test does not tell us what is the true distribution of the data in the population as a whole. We may consider a range of options to deal with this uncertainty regarding the underlying distribution of the data.

One approach is to fit a model using an over-dispersed Poisson regression. This remains a model-based approach, and continues to assume that the Poisson model is correct, but accounts for the over-dispersion. The model includes a multiplier to the Poisson variance function, which is assumed to be constant for all data. The multiplier is estimated using a quasi-likelihood, where the Poisson model is fit and the constant is estimated *post hoc* and used to inflate the standard error estimates. The results of the over-dispersed Poisson are shown in column 2 of Table 15. (Column repeats the full model from Table 14 for comparison.) The parameter estimates do not change, but the standard errors are larger, as the model must account for the dispersion factor in its estimate. The coefficient on Ln(Septics) remains statistically significant at the 1 percent level, and the coefficient on the soil drainage parameter is significant at the 5 percent level. The housing density parameter is no longer statistically different from zero, and the Sewered dummy and Sewered-Ln(Septic) interaction term are not jointly significant. The USEPA Regional dummies remain jointly significant.

Another approach is to relax the assumption that the Poisson model is correct; we continue to estimate the model using a Poisson model, but make no assumption regarding the underlying data. This increases the standard errors of the estimate because we introduce an additional source of uncertainty into the model – the uncertainty about the form of the model. The coefficients are estimated assuming the Poisson model holds for both the mean and the variance, and the standard errors are adjusted to allow for possible violations of the variance assumption. In other words, the estimates do not rely on a specific functional form for the variance of LCSSs. The standard errors are robust because they are consistent even if the data violate the assumptions used to produce estimates of the regression coefficients. We consider this a model assisted approach, and its results are shown in column 3 of Table 15.

These approaches ignore the sampling design. The next two approaches incorporate information about the sampling design by weighting the data by the inverse of the selection probabilities. The fourth approach, shown in column 4, fits an over-dispersed Poisson, weighted by the selection probabilities. The fifth approach estimates a weighted model with robust standard errors. The weighting scheme changes the parameter estimates, and increases the standard errors. The Ln(Septic) coefficient remains significant in the over-dispersed model, as do the housing density and soil drainage parameters. Only the Ln(Septic) parameter remains significant in the model with robust standard errors. The USEPA Regional dummy variables are jointly significant in both weighted models.

As we will see in the discussion of the estimates of the total, the changes in the parameter estimates do not dramatically affect the estimate of the total number of wells. The weighted model's changes in the robust standard errors greatly increase the mean square error of the estimated total.

Table 15. Comparison of Standard Poisson,
Over-Dispersed Poisson, and Robust Standard Error Models

	(1)	(2)	(3)	(4)	(5)
Ln(Septics)	0.849 (0.044)**	0.849 (0.100)**	0.849 (0.106)**	0.719 (0.089)**	0.719 (0.253)**
90-100% Sewered	1.320 (1.131)	1.320 (2.549)	1.320 (1.537)	0.621 (2.858)	0.621 (2.711)
Sewered*Ln(Septic)	-0.112 (0.236)	-0.112 (0.532)	-0.112 (0.327)	-0.098 (0.591)	-0.098 (0.579)
Total Housing Units/Sq. Mile	-0.001 (0.000)**	-0.001 (0.001)	-0.001 (0.001)	-0.002 (0.001)*	-0.002 (0.004)
% Area Poorly Drained	-0.010 (0.002)**	-0.010 (0.004)*	-0.010 (0.005)*	-0.011 (0.004)**	-0.011 (0.009)
USEPA Region 2	1.018 (0.125)**	1.018 (0.282)**	1.018 (0.354)**	1.273 (0.327)**	1.273 0.697)*
USEPA Region 3	-0.007 (0.114)	-0.007 (0.257)	-0.007 (0.189)	0.207 (0.293)	0.207 (0.381)
USEPA Region 4	0.546 (0.103)**	0.546 (0.232)**	0.546 (0.185)**	0.653 (0.270)**	0.653 (0.343)*
USEPA Region 5	-0.424 (0.127)**	-0.424 (0.286)	-0.424 (0.206)*	-0.410 (0.291)	-0.410 (0.347)
USEPA Region 6	0.519 (0.136)**	0.519 (0.307)*	0.519 (0.339)	0.999 (0.332)**	0.999 (0.711)
USEPA Region 7	-0.413 (0.197)*	-0.413 (0.445)	-0.413 (0.303)	-0.589 (0.446)	0.589 (0.508)
USEPA Region 8	0.960 (0.112)**	0.960 (0.252)**	0.960 (0.257)**	0.919 (0.331)**	0.919 (0.766)
USEPA Region 9	0.417 (0.125)**	0.417 (0.281)	0.417 (0.202)*	0.188 (0.393)	0.188 (1.275)
USEPA Region 10	-0.171 (0.164)	-0.171 (0.370)	-0.171 (0.315)	0.020 (0.749)	0.020 (2.168)
Constant	-3.143 (0.337)**	-3.143 (0.760)**	-3.143 (0.811)**	-2.235 (0.687)**	-2.235 (1.851)
Observations	99	99	99	99	99

Standard errors are in parentheses.
* significant at 5% level; ** significant at 1% level.

In addition to the sampling weights, the model could incorporate information about the clustering in the sample. We do not believe the data support the use of this information. The strata do not appear to be explain much of the variation in the data, so any adjustment would be small. Furthermore, the method used to estimate the standard errors that incorporate the information about the clustering can be highly variable when the number of clusters drawn from each strata is small (Kott, 1994).

To assess the potential impact of the clustering of the data on the estimated standard errors, we estimated the intra-class correlation, which measures the similarity of data from tracts within a cluster relative to the variability of the data across all clusters (Snedecor and Cochran, 1980). To control for the characteristics of the tract that are associated with the number of LCSSs, we conducted a one-way analysis of variance of the residuals from the Poisson model. While this linear approximation of the intra-class correlation is biased, it gives an indication of the effect of the clusters on the estimated standard error. This measure of the intra-class correlation is approximately 1 percent; therefore, we concluded that the potential impact of the correlation among tracts within clusters on the estimated standard errors is small.

A.1.2 Estimating the Total Number of LCSSs and Calculating the Standard Error of the Estimated Total

To get the total number of LCSSs in the nation, we sum the estimates across each census tract. The three models that are not weighted by the inverse of the selection probabilities produce the same estimated total, but their standard errors differ. The two weighted models each estimate the same, slightly higher total, again with different standard errors. Assume Y_i is the number of LCSSs in tract i, which has an over-dispersed Poisson distribution with mean $:_i$, variance $V(:_i) \times N = N\,:_i$, and also that $:_i = \exp(\mathbf{x_i}^T\$)$. N is a dispersion factor. The row vector $\mathbf{x_i}^T$ is the set of explanatory variables for tract i, that may or may not be in the sample, and $\$$ is the column vector of regression parameters. The estimated total for the entire country is:

$$(9) \qquad T = \sum_{i=1}^{s} Y_i + \sum_{i=s+1}^{N} \exp(\mathbf{x_i}^{T}\,\hat{\ })$$

assuming that the tracts are numbered so that tracts 1 to s are the sampled tracts and tracts s+1 to N are the non-sampled tracts. In this model-based approach we are assuming that the LCS values for the N tracts are independent.

Since we are predicting the random total for the entire country, we need the prediction error variance:

$$(10) \qquad V(T) = E\left\{\left(T - \sum_{i=1}^{N} Y_i\right)^2\right\} = E\left\{\left\{\sum_{i=s+1}^{N} (Y_i - \exp(\mathbf{x_i}^{T}\hat{b})\right\}^2\right\}$$

where V(T) is the variance of the estimated total T. Expand the expression inside the parentheses as a Taylor series about the true regression parameter vector $\$$.

$$\sum_{i=s+1}^{N}\left\{Y_i - \exp(\mathbf{x_i}^{T}\hat{b})\right\} \approx \sum_{i=s+1}^{N}\{Y_i - \exp(\mathbf{x_i}^{T}b)\} - \exp(\mathbf{x_i}^{T}b)\sum_{j} x_{ij}(\hat{b}_j - b_j)$$

$$(11) \qquad = \sum_{i=s+1}^{N}\left\{Y_i - m_i\right\} - \sum_{i=s+1}^{N}\sum_{j} m_i x_{ij}(\hat{b}_j - b_j)$$

$$= \sum_{i=s+1}^{N}(Y_i - m_i) - A = B$$

where x_{ij} is the jth explanatory variable for the ith tract.

The mean of the last expression (B) is zero. Therefore the expected value of B squared equals the variance of B. But B is the sum of N-s independent random variables Y_i ! :$_i$ minus A (the double sum). Since A is a function of the first s values (the sampled tracts), this double sum is independent of the other N-s terms. Therefore:

$$V(T) = Var(B) = \sum_{i=s+1}^{N} V(Y_i) + V(A)$$

(12) $$= \sum_{i=s+1}^{N} f m_i + V(\sum_{j} K_j \hat{b}_j)$$

$$= \sum_{i=s+1}^{N} f m_i + \sum_{j} K_j^2 V(\hat{b}_j) + 2\sum_{j<k} K_j K_k Cov(\hat{b}_j, \hat{b}_k)$$

where:

(13) $$K_j = \sum_{i=s+1}^{N} m_i x_{ij}$$

Thus to estimate the mean square error of prediction, substitute estimates for the dispersion parameter, the mean values for the non-sampled tracts, and the variances and covariances of the regression parameters.

$$Estimated\ V(T) =$$

(14) $$\sum_{i=s+1}^{N} \hat{f}\hat{m}_i + \sum_{j} \hat{K}_j^2 V(\hat{b}_j) + 2\sum_{j<k} \hat{K}_j \hat{K}_k Cov(\hat{b}_j, \hat{b}_k)$$

where:

(15) $$\hat{K}_j = \sum_{i=s+1}^{N} \hat{m}_i x_{ij}$$

An approximate 95 percent prediction interval for the country total is the estimated total plus or minus 1.96 times the square root of the estimated V(T).

Table 16 shows each model's estimated total, standard error, and 95 percent prediction interval.

A.2 Storm Water Drainage Wells

The occurrence of storm water drainage wells in the sample was a relatively rare event: 22 of the 99 tracts contained storm water drainage wells. Among the tracts in the sample that contained wells, the average number of wells is 19. The median number of wells is 2, and most tracts contain fewer than 6 wells. The high average is due two outliers: one tract that contains 81 wells and a second that contains 210. If these two tracts are excluded, the average number drops to 6.

Table 16. Models' Estimates of the Number of LCSSs in the Eligible Tracts in the United States

	Estimated Number of LCSSs	Standard Error	95% Confidence Interval	
			Lower Boundary	Upper Boundary
Poisson Model	289,385	8,411	272,899	305,871
Over-Dispersed Poisson Model	289,385	18,806	252,526	326,244
Poisson Model with Robust Standard Errors	289,385	21,165	247,902	330,868
Weighted Over-Dispersed Poisson Model	303,169	18,905	266,114	340,224
Weighted Poisson Model with Robust Standard Errors	303,169	46,973	211,101	395,237

To account for the large number of tracts that do not contain storm water drainage wells, we used a two-part model (Duan et al., 1983). Formally, the expectation of the number of wells in a given tract i is expressed by:

(16) $E[SDW_j] = P(SDW_j>0) * E[SDW]|_{SDW>0}$

Where: $E[SDW_i]$ = The expectation of the number of storm water drainage wells in tract i.

$P(SDW_j>0)$ = The probability that tract i contains storm water drainage wells.

$E[SDW]|_{SDW>0}$ = is the expectation of the number of the number of storm water drainage wells in tracts that contain wells.

The likelihood function of the two-part model can be separated into two components. The first part, which provides an estimate of the probability that a tract contains storm water drainage wells, is estimated using a probit model. The second part, which provides an estimate of the number of wells in tracts that contain wells, is a weighted average of the number of wells in tracts in the sample that contain wells.

Probit Model of Probability that a Tract Contains Storm Water Drainage Wells

The probability that a tract contains a well is estimated using a probit regression (Aldrich and Nelson, 1984). We estimate the probability as a function of the density of housing built in the tract before 1970, the percentage of the area with poorly drained soils, and mean annual precipitation. The estimated probability is given by:

$$P(SDW > 0) = \Phi\left(b_0 + b_1 Density_{70} + b_2 Drainage + b_3 MAP\right) \quad (17)$$

Where: M = the standard normal cumulative distribution function;
$Density_{70}$ is the number of housing units built before 1970, per square mile;

Drainage is the percentage of the tract with poor soil drainage; and

MAP is the mean annual precipitation in the tract, measured in inches.

The distributions of the density of housing built before 1970 and mean annual precipitation are shown in Figures 9 and 10, respectively. (The distribution of the percentage of area with poorly drained soil is shown in Figure 1.4 above.)

Figure 9. Distribution of Density of Housing Built Prior to 1970 in Sample of 99 Census Tracts

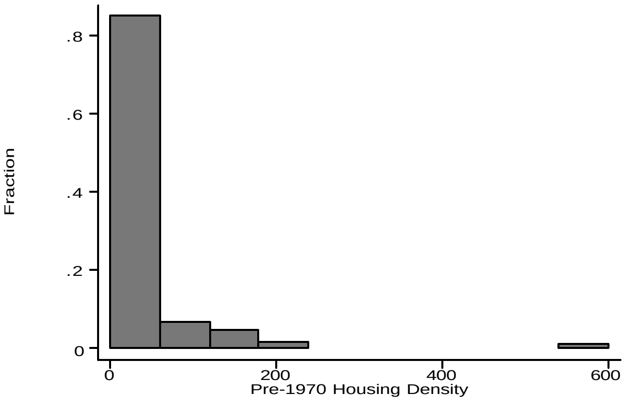
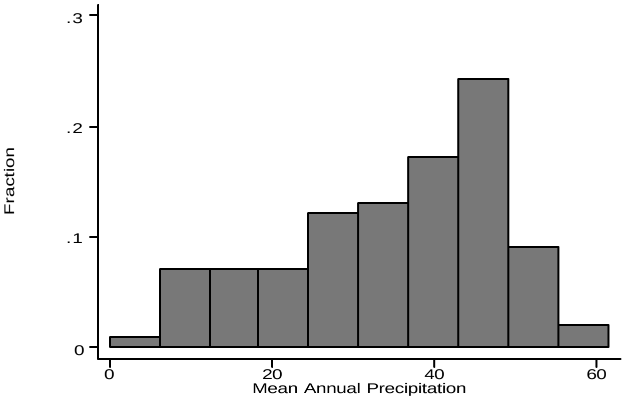


Figure 10. Distribution of Mean Annual Precipitation in Sample of 99 Census Tracts



The results of this model are shown in the first column of Table 17. Rather than show the parameters, the table shows the change in the probability for a change in each explanatory variable, evaluated at the mean of the data (dF/dX). Standard errors are shown in parentheses.

We assumed that tracts with greater densities of older housing are more likely to contain storm drainage wells. This is based on the assumption that development that occurred 20 to 30 years or longer ago was more likely to use these wells than more recent development, and that the number of wells will increase as housing density increases. (The year 1970 was used because the 1990 Census identifies houses built before and after that year.) The coefficient on this density parameter is statistically significant at the 5 percent level. An alternative specification used total housing density, which was not statistically significant.

Table 17. Probit Models of the Occurrence of Storm Water Drainage Wells

	(1)	(2)
Density_{70}: Pre-1970 housing density	0.001 (0.001)*	0.001 (0.000)**
Drainage: % Area Poorly Drained	-0.008 (0.003)*	-0.010 (0.003)**
MAP: Mean Annual Precipitation	-0.007 (0.003)*	0.005 (0.006)
South [a]		-0.173 (0.065)* †
Mid-West [a]		-0.121 (0.081)†
West [a]		0.400 (0.284)†
Observed probability	0.222	0.222
Predicted probability [b]	0.181	0.142
Observations	99	99

Standard errors in parentheses. Reference region is the Northeast.

[a] dF/dx is for a discrete change of dummy variable from 0 to 1.

[b] At mean of the independent variables.

* Significant at 5% level; ** significant at 1% level; † jointly significant at 5% level.

Soil drainage is inversely related to the probability of the occurrence of storm water drainage wells. This is consistent with the assumption that storm water drainage wells are not adequate sources of drainage in areas with poor soil drainage. Mean annual precipitation also is inversely related, which implies that areas with large amounts of rain or snow fall must rely on other sources of drainage. This conclusion is supported by some of the more qualitative data collected during the site visits. Both the soil drainage and mean annual precipitation coefficients are statistically significant.

To account for possible differences by region, we considered a model that includes three dummy variables that indicate tracts in the South, Mid-West, and West. The reference group is the Northeast. (We used these larger regional categories rather than the USEPA Regions because so few tracts contain wells.) This model is shown in column 2 of Table 17. The dummy variables are jointly significant at the 5 percent level, indicating that the probability of the occurrence of storm water drainage wells does vary across region. The probability is higher in the West than the rest of the country, and the lowest in the South. The regional dummy variables are correlated with the mean annual precipitation variable, which becomes statistically insignificant when we introduce the regional dummy variables. In fact, the differences in precipitation may explain much of the observed regional differences, so we will use the model in column 1.

We also considered other explanatory variables, which did not prove to be significant. These included the percentage of the area with susceptible bedrock, the percentage of the area with sand or gravel soil, the percentage of the housing units that are mobile homes, the percentage of structures that contain five or more housing units, and the percentage of housing units that are part-time residences.

We also considered transforming the poorly drained soil and housing density variables. The model with these transformed variables did not perform as well as the model with the untransformed variables.

This specification of the model does not incorporate information about the sampling design. Table 18 compares this model to one that weights the data by the inverse of the selection probabilities. The unweighted model is shown in column 1, and the weighted model is in column 2.

Table 18. Comparison of Weighted and Unweighted Probit Models of the Occurrence of Storm Water Drainage Wells

	(1)	(2)
	Unweighted Probit	Weighted Probit[a]
Density$_{70}$: Pre-1970 housing density	0.001	0.001
	(0.001)*	(0.001)*
Drainage: % Area Poorly Drained	-0.008	-0.007
	(0.003)*	(0.003)**
MAP: Mean Annual Precipitation	-0.007	-0.006
	(0.003)*	(0.004)
Observed probability	0.222	0.196
Predicted probability [b]	0.181	0.152
Observations	99	99

Standard errors are in parentheses.
[a] Weighted by inverse of selection probability.
[b] At mean of the independent variables.
* significant at 5% level; ** significant at 1% level

The weights result in modest changes to the parameters. Because the coefficient on mean annual precipitation declines slightly and the standard error increases, it is not significant in the weighted model.

While each of the parameters is statistically significant in the unweighted model, the probit model explains relatively little of the variation in the probability that a tract will contain storm water drainage wells. The pseudo R^2 for the model is just over 12 percent, which implies that the model explains just less than **C** of the total variation. This is due, in part, to the relatively small size of our sample. As we will see, this contributes to a relatively large variance in the estimate of the total number of wells.

Estimating the Average Number of Storm Water Drainage Wells in Tracts with Wells

The probit model provides an estimate of the probability that storm water wells are used in a given tract; the expectation of the number of wells is then equal to that probability multiplied by the average number of storm water drainage wells in tracts containing wells. We compute the average number wells in tracts with wells, weighted by the inverse of the selection probabilities to account for the stratification of our sample. (Weights are used because the simple mean does not include meaningful predictors, unlike the probit model discussed above.) We also distinguish tracts in which more than 10 percent of the soil is sand and gravel from other tracts. The site visits indicated that the occurrence of sand and gravel soil is conducive to the use of storm water drainage wells. The average number of wells in tracts in which more than 10 percent of the soil is sand or gravel is 18.2; the average for the rest of the tracts in the sample is 6.9. There is considerable noise around the estimated means, and the difference is not statistically significant.

A.2.1 Estimating the Total Number of Storm Water Drainage Systems and Calculating the Standard Error of the Estimated Total

The model's prediction of the expectation of the number of wells in tract i, given the tract's characteristics, is given by equation 18:

$$E[SDW] = \begin{cases} \Phi\left(b_0 + b_1 Density_{70} + b_2 Drainage + b_3 MAP\right) * \overline{SDW}|_{\mathrm{SDW>0\ and\ PSG>10}} \ ; \text{ if psg} > 10\% \\ \Phi\left(b_0 + b_1 Density_{70} + b_2 Drainage + b_3 MAP\right) * \overline{SDW}|_{\mathrm{SDW>0\ and\ PSG\leq 10}} \ ; \text{ if psg} \leq 10\% \end{cases} \quad (18)$$

Where: M is the standard normal cumulative distribution function;

Density_{70} is the number of housing units built before 1970, per square mile;

Drainage is the percentage of the tract with poor soil drainage;

MAP is the mean annual precipitation in the tract, measured in inches;

PSG is the percentage of area with sand or gravel soil;

$\overline{\mathrm{SDW}}|_{\mathrm{SDW>0\ and\ PSG>10}}$ is the average number of storm water drainage wells in tracts with wells and in which greater than 10 percent of the soil is sand or gravel; and

$\overline{\mathrm{SDW}}|_{\mathrm{SDW>0\ and\ PSG\leq 10}}$ is the average number of storm water drainage wells in tracts with wells and in which less than 10 percent of the soil is sand or gravel.

The total number of wells in non-urbanized areas is equal to the sum across each eligible tract in the nation of the estimated number of wells. The error structure of this two-part model is complex, and is difficult to derive analytically. Therefore, we use a bootstrap to estimate the mean square error of the estimated total (Effron and Tibshirani, 1993). The bootstrap draws 1,000 samples of 99 tracts with replacement from our sample, and estimates the parameters of the two-part models specified in Table 18 for each sample. Using these parameters, it then predicts the total number of wells in the non-urbanized areas of the country, creating 1,000 estimated totals. The standard error of the total is equal to the standard error of this bootstrap sample. The 95 percent prediction interval is equal to the 2.5th and 97.5th percentiles. Table 19 shows the estimated total, standard error, and 95 percent prediction interval for the weighted and unweighted models.

As with the LCSS model, we chose not to incorporate information about the clustering of the sample. Using an approach similar to the one for LCSSs, we estimated the intra-class correlation, which was less than 1 percent. Therefore, we assumes the potential impact on the estimated standard errors of the model is small.

Table 19. Unweighted and Weighted Two-Part Model Estimate of the Number of Storm Water Drainage Wells in Non-Urban Areas in the United States [a]

			95% Confidence Interval	
	Estimated Number of LCSSs	Standard Error	Lower Boundary	Upper Boundary
Unweighted model	63,998	35,278	13,409	145,174
Weighted model	59,105	34,107	11,345	131,600

[a] Includes eligible tracts, plus tracts in densely populated areas surrounding urban areas.

ATTACHMENT B
PROBABILITY OF SELECTION INTO THE SAMPLE

Section 2 of this Appendix describes in detail how the sample of census tracts was developed. The probability of a tract being selected for the sample depends on the state the tract is in, the strata it is assigned to given its characteristics, and the geographic cluster it is assigned to. This attachment describes how we estimate these selection probabilities.

Each target has two opportunities to be included in the sample: a tract can be selected in the first stage as a target tract, or it can be selected from among the tracts surrounding these targets in the second stage. Formally, if we call the probability of selection in stage 1 P(A) and the probability of selection in stage 2 P(B), the probability of selection is given by:

(19) $P(S) = P(A \text{ or } B) = P(A) + P(B) – P(A \text{ and } B)$

Because A and B are mutually exclusive events (a tract can be selected in stage 1 or stage 2, but not both), P(A and B) is equal to zero. Therefore, the selection probability is:

(20) $P(A \text{ or } B) = P(A) + P(B)$

P(B), or the probability that a tract is selected in the second stage, is itself the joint probability of two events: the probability of a tract being within 100 miles of a target tract and the probability that the tract is selected from among all the tracts within 100 miles of a target. If we call the former P(C) and the latter P(D), we get:

(21) $P(B) = P(C \text{ and } D)$

(22) $P(B)=P(C)*P(D|C)$

Substituting back into (20), we get:

(23) $P(S) = P(A) + P(C)*P(D|C)$

The remainder of the is document describes how each probability is calculated.

B.1 Calculate Stage 1 Probability [P(A)]

First, define some terms:

N_S = Total number of eligible tracts in State S, for S=1 to 48 (Alaska and Hawaii are not included in the population).

N_R = Number of tracts in stratum R, for R=1 to 16.

T_R = Number of targets drawn from stratum R, for R=1 to 16.

The number of tracts varies greatly by state, so some states have a much higher probability of containing a tract that is included in the sample than others. To include as many states as possible in the sample, we weighted the data so each state has the same chance of being included in the sample. The weight for each observation i in State S, which we call W_{1i}, is equal to:

(24)
$$W_{1i} = \frac{\left(\frac{1}{48}\right)}{\left(\frac{N_s}{\sum_{S=1}^{48} N_S}\right)}$$

Using the addition rule with mutually exclusive events, the selection probability in the first stage, P(A), is given by:

(25)
$$P(A) = T_R * \left(\frac{w_{1i}}{\sum_{i=1}^{N_R} w_{1i}} \right)$$

Equation 25 is an approximation. In a small number of cases, two target tracts were very close to each other. Because the list of tracts surrounding the two potential targets tracts were virtually identical in these cases, one of the two tracts was randomly dropped. In other cases, there was some overlap in the areas covered by two targets. In these cases, the surrounding tracts were randomly assigned to one target. It is believed that the effect of these adjustments is very small, and it is not included in equation 25.

B.2 Calculate Conditional Probability of Selection in Stage 2 if within 100 Miles of a Target Tract [P(D|C)]

First, some further notation:

U_R = Number of tracts in stratum R surrounding all the targets, for R=1 to 16.

S_R = The number of tracts to draw from stratum R in the second stage, for R=1 to 16.

U_t = The number of tracts surrounding target t.

Each tract is weighted to reflect the number of tracts we wish to draw from each stratum. The weights are equal to:

(26) $$W_{2i} = \frac{\left(S_R \Big/ \sum_{R=1}^{16} S_R \right)}{\left(U_R \Big/ \sum_{R=1}^{16} U_R \right)}$$

We select two tracts from each cluster. Using the addition rule again, the probability that a tract is selected from the among all the tracts within 100 miles of a given target is given by:

(27) $$P(D \mid C) = 2 * \left(\frac{w_{2i}}{\sum_{i=1}^{U_t} (w_{2i})} \right)$$

Equation 27 is an approximation of the conditional probability of selection in the second stage. Each tract is within 100 miles of multiple potential targets, so the probability of a tract being selected in this stage depends on which target is selected. Two factors can affect this probability: the number of tracts surrounding each possible target, and the distribution of those tracts across the sixteen strata. There is considerable overlap among the potential targets, because they all are within 100 miles of the tract in question. Therefore, we expect the variation in both the number of tracts surrounding each target and the distribution across the strata to be relatively small, and we believe equation 27 is a reasonable approximation of the conditional probability of selection in the second stage.

B.3 Calculate the Probability that Tract is Not a Target and is within 100 Miles of a Target [P(C)]

For a tract to be eligible for selection in the second round, two things must occur. First, it must not be selected in the first round. Second, it must be within 100 miles of a target tract that was selected in the first round. The probability that a tract is not selected in the first round is equal to 1-P(A). The probability that a tract is within 100 miles of a target is the equivalent of the probability that a tract within 100 miles of a given tract is selected as a target. Some final notation:

G_R = The number of tracts in stratum R surrounding a given tract, for R=1 to 16.

The probability of a given tract being within 100 miles of a selected target is given by the following joint probability:

$$(28) \qquad 1-\prod_{R=1}^{16}\left(\prod_{j=0}^{T_R-1}\left(\frac{\sum_{i=1}^{N_R} W_{1i}-\sum_{i=1}^{G_R} W_{1i}-j}{\sum_{i=1}^{N_R} W_{1i}-j}\right)\right)$$

This uses the law of total probability. For each stratum, we calculate the joint probability that each target selected is not from a tract within 100 miles of a given tract. This is equal to the product of one of these other tracts being selected the first, multiplied by the probability that one is selected in the second conditional on one being selected in the first, round, and so on. We then multiply these products across the sixteen stratum, which equals the probability that all the targets are drawn from tracts that are not within 100 miles of the given tract. The probability that at least one target is drawn from the tracts within 100 miles of a given tract is equal to 1 minus this probability.

The probability that tract makes into the second round, P(C), is given by the product of the probability it is not selected in the first round, and the probability that it is within 100 miles of a target tract:

$$(29) \qquad P(C)=(1-P(A))*\left(1-\prod_{R=1}^{16}\left(\prod_{j=0}^{T_R-1}\left(\frac{\sum_{i=1}^{N_R} W_{1i}-\sum_{i=1}^{G_R} W_{1i}-j}{\sum_{i=1}^{N_R} W_{1i}-j}\right)\right)\right)$$

ATTACHMENT C
CHARTS OF GEOLOGIC AND DEMOGRAPHIC CHARACTERISTICS OF ELIGIBLE TRACTS AND TRACTS IN SAMPLE

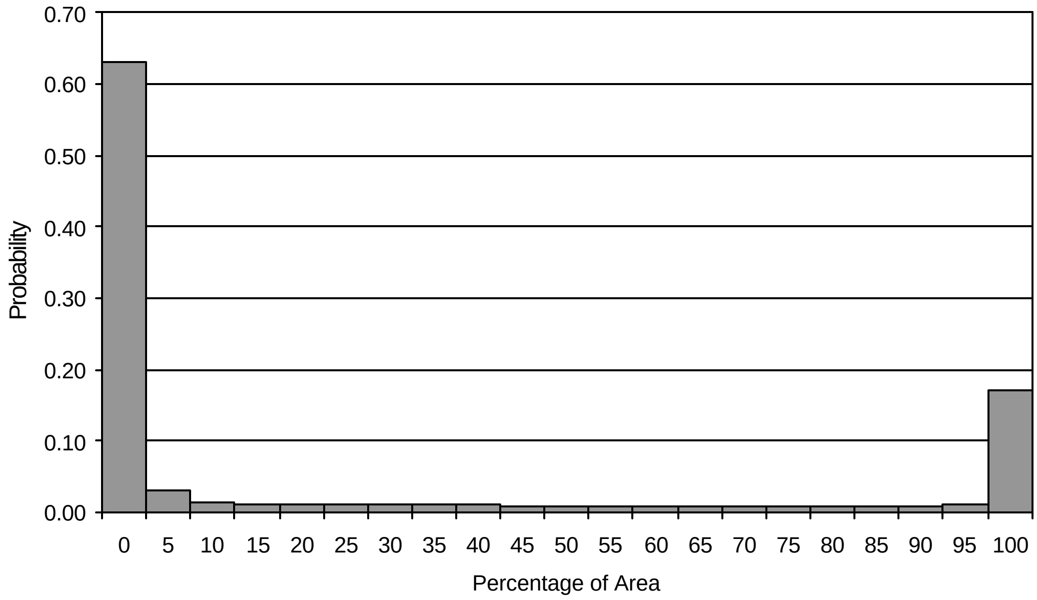
The following charts show the distribution of several of the geologic and demographic characteristics of the census tracts. Two graphs are shown in each figure. The first shows the distribution of the characteristic for the 18,578 eligible tracts in the sample frame; the second shows the distribution for the 99 tracts in the sample.

The first set of figures show the distribution of the characteristics used to stratify the data. Figure 11 shows the distribution of the percentage of area with susceptible bedrock for the 18,578 eligible tracts and the 99 tracts in the sample. Figure 12 shows the distribution of the percentage of area with sand or gravel soil. Figure 13 shows the distribution of the percentage of area on public sewers. Figures 14 through 16 show the distribution of the percentage of structures containing five or more housing units, the percentage of housing units that are mobile homes, and the percentage of housing units that are part-time residences. Figure 17 shows the distribution of the percentage of the area with bedrock within five feet of the surface.

The remaining figures show the distribution of characteristics used in the inventory model as explanatory variables. Figure 18 shows the distribution of the number of households per square mile, and Figure 19 shows the distribution of the number of households built prior to 1970 per square mile. Figure 20 shows the percentage of area with poorly drained soils. Figure 21 shows mean annual precipitation.

Figure 11. Percentage of Area with Susceptible Bedrock

Eligible Tracts



Sample Tracts

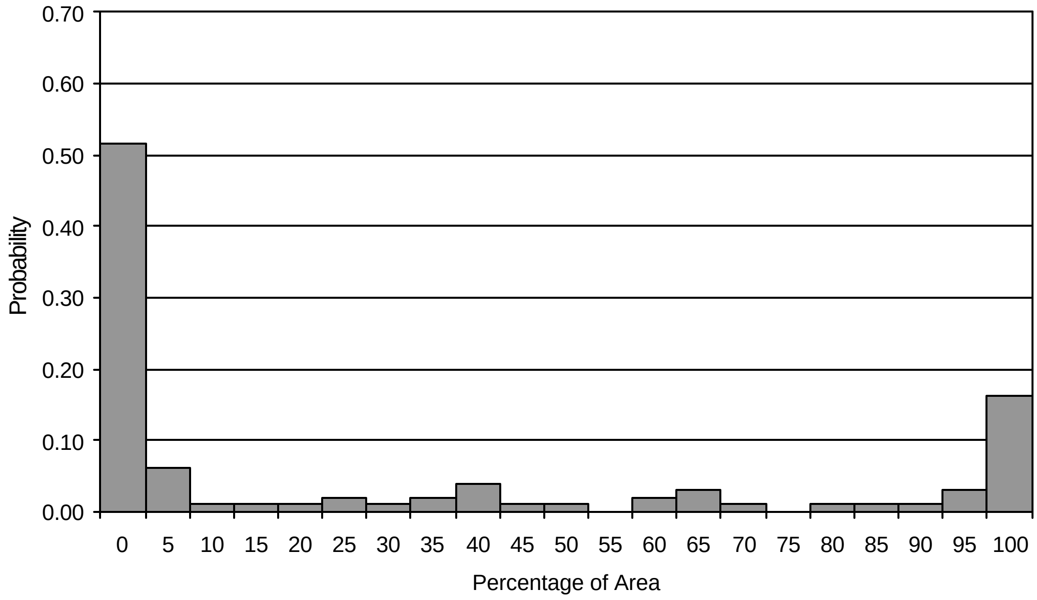
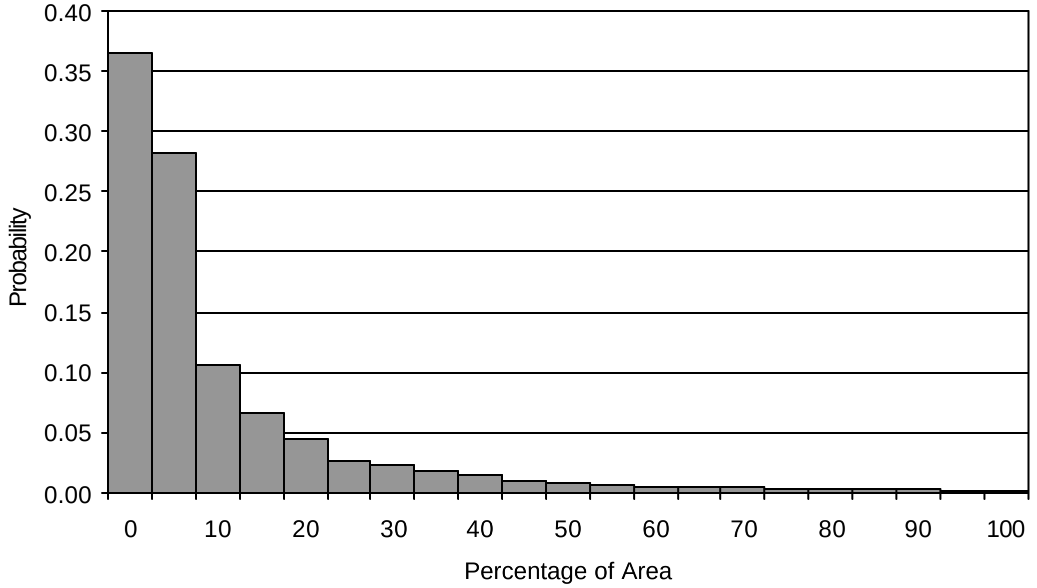


Figure 12. Percentage of Area with Sand and Gravel Soil

Eligible Tracts



Sample Tracts

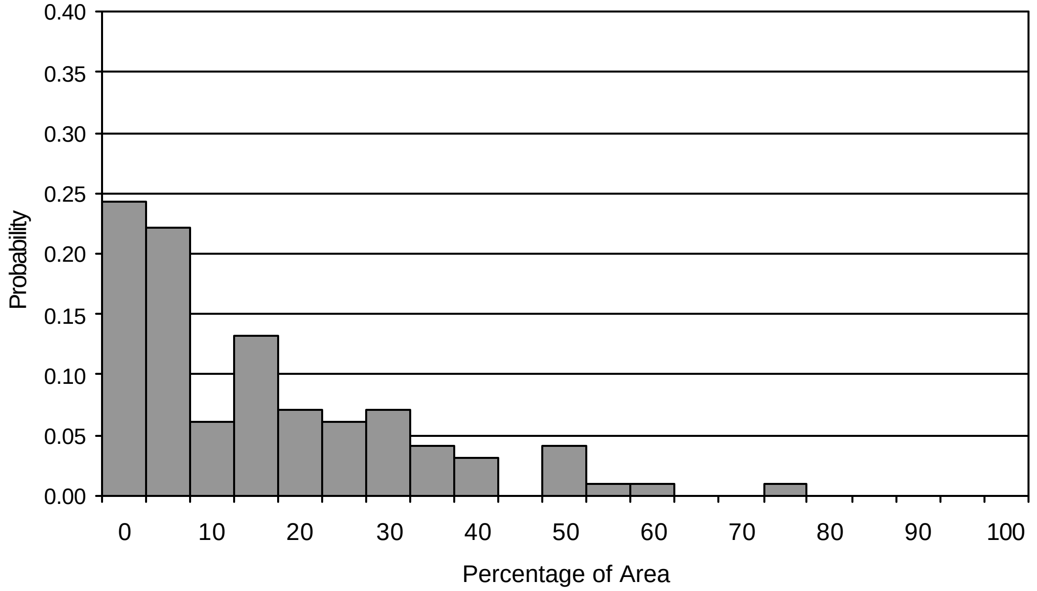
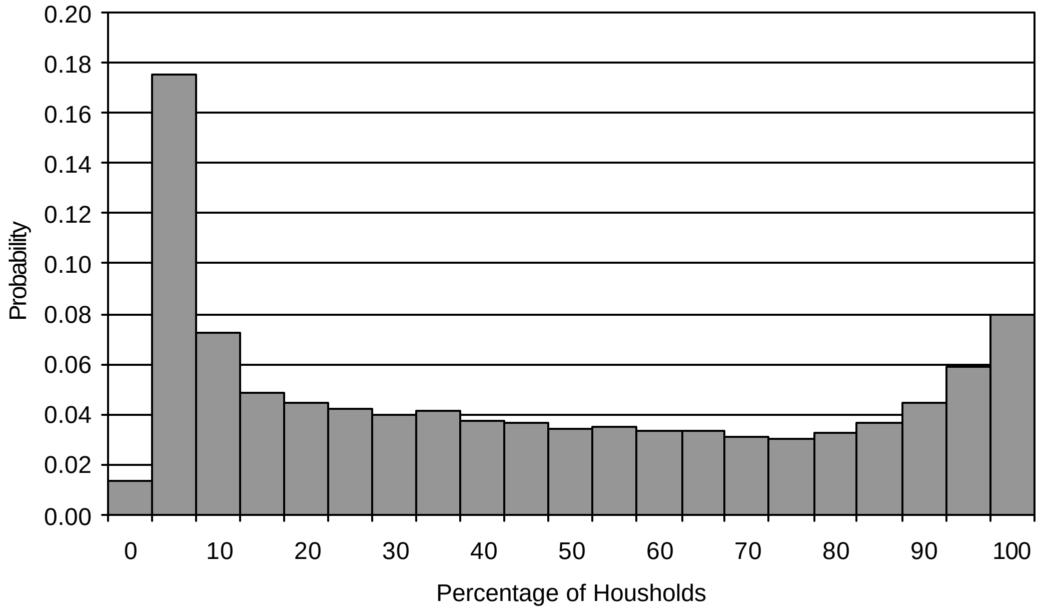


Figure 13. Percentage of Households on Public Sewers

Eligible Tracts



Sample Tracts

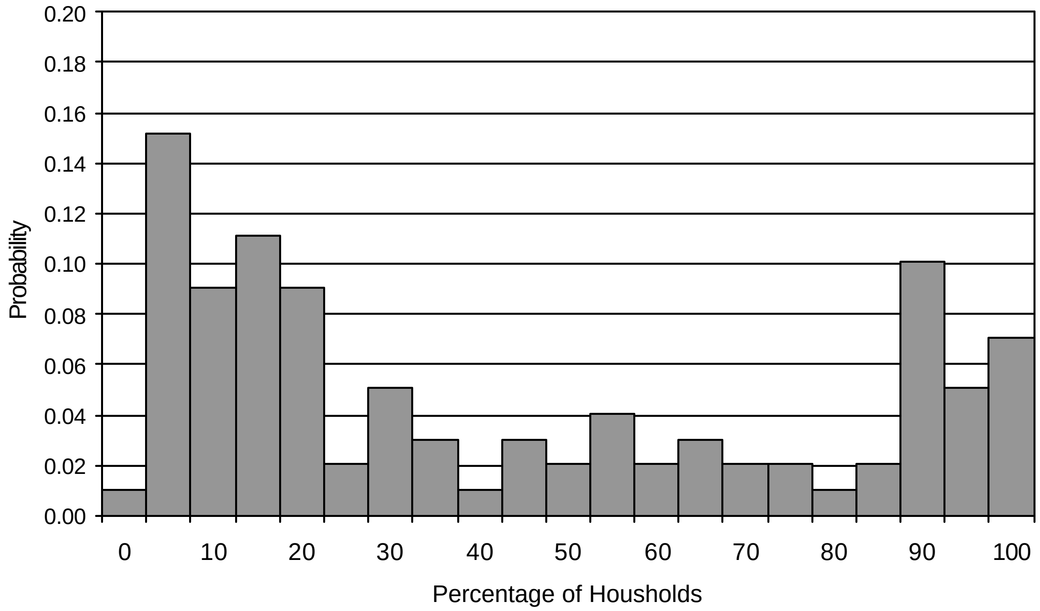
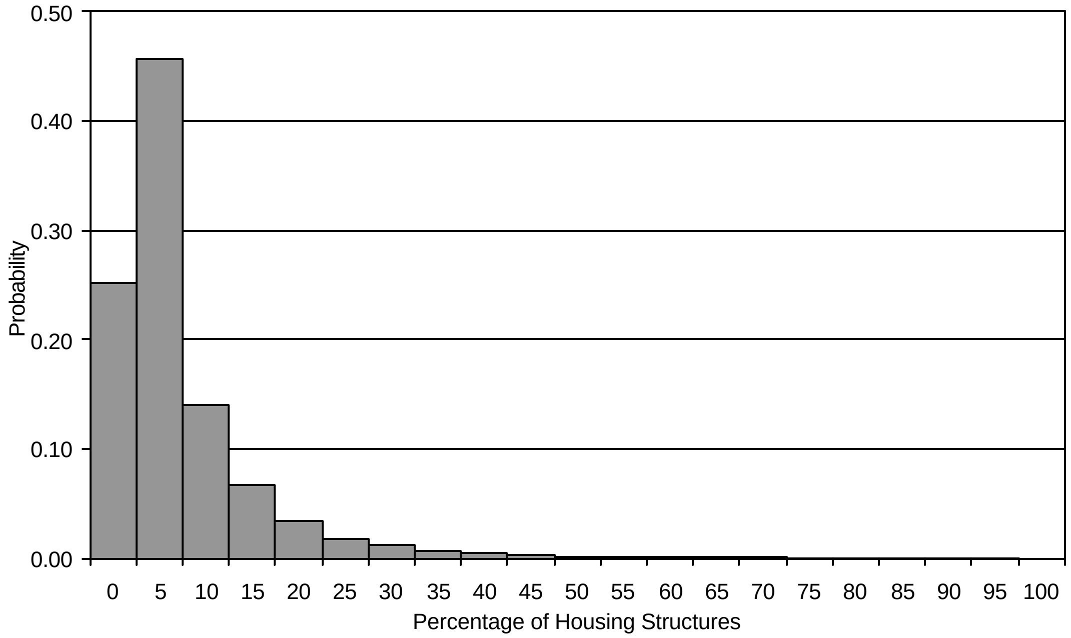


Figure 14. Percentage of Structures with 5 or More Housing Units

Eligible Tracts



Sample Tracts

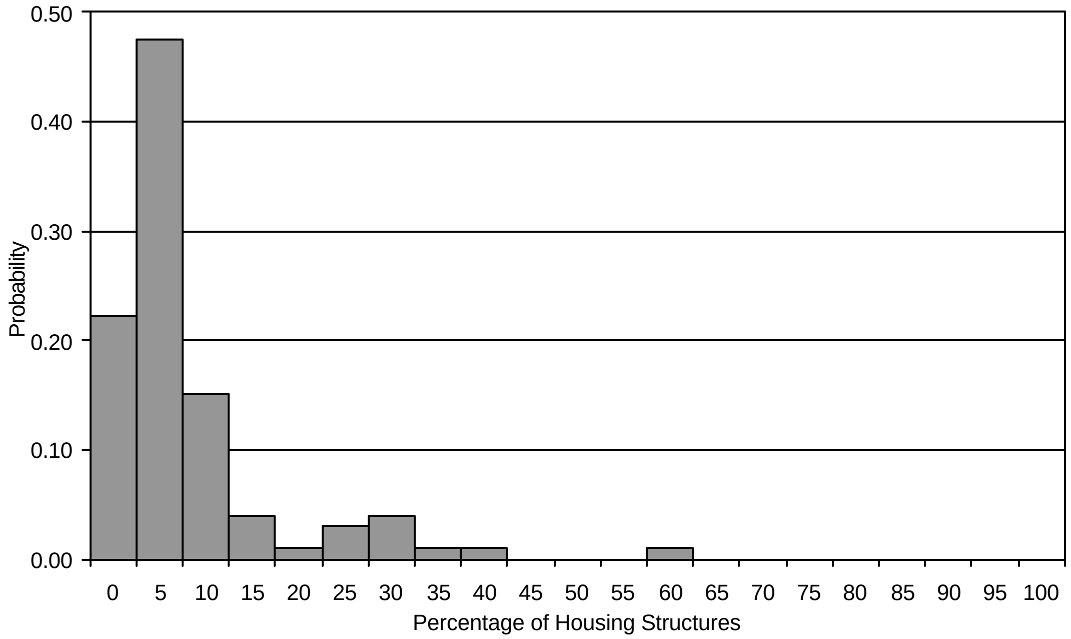
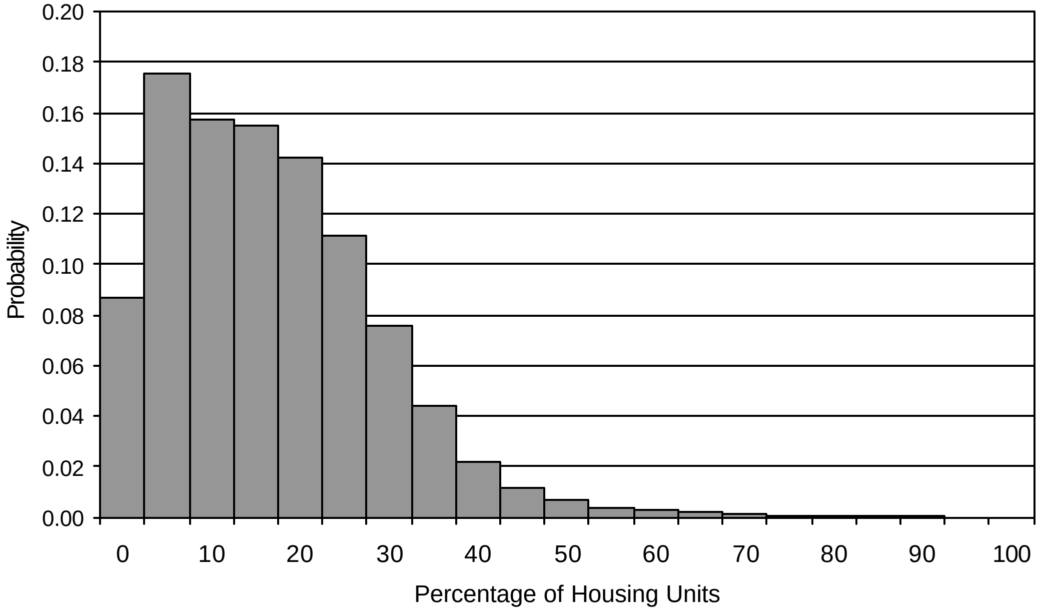


Figure 15. Percentage of Housing Units that Are Mobile Homes

Eligible Tracts



Sample Tracts

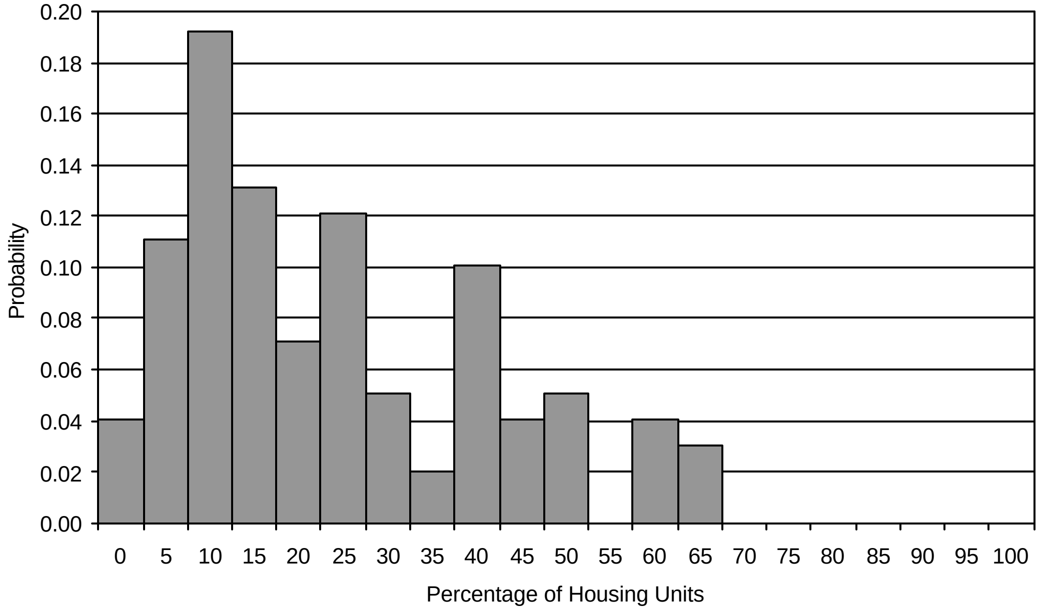
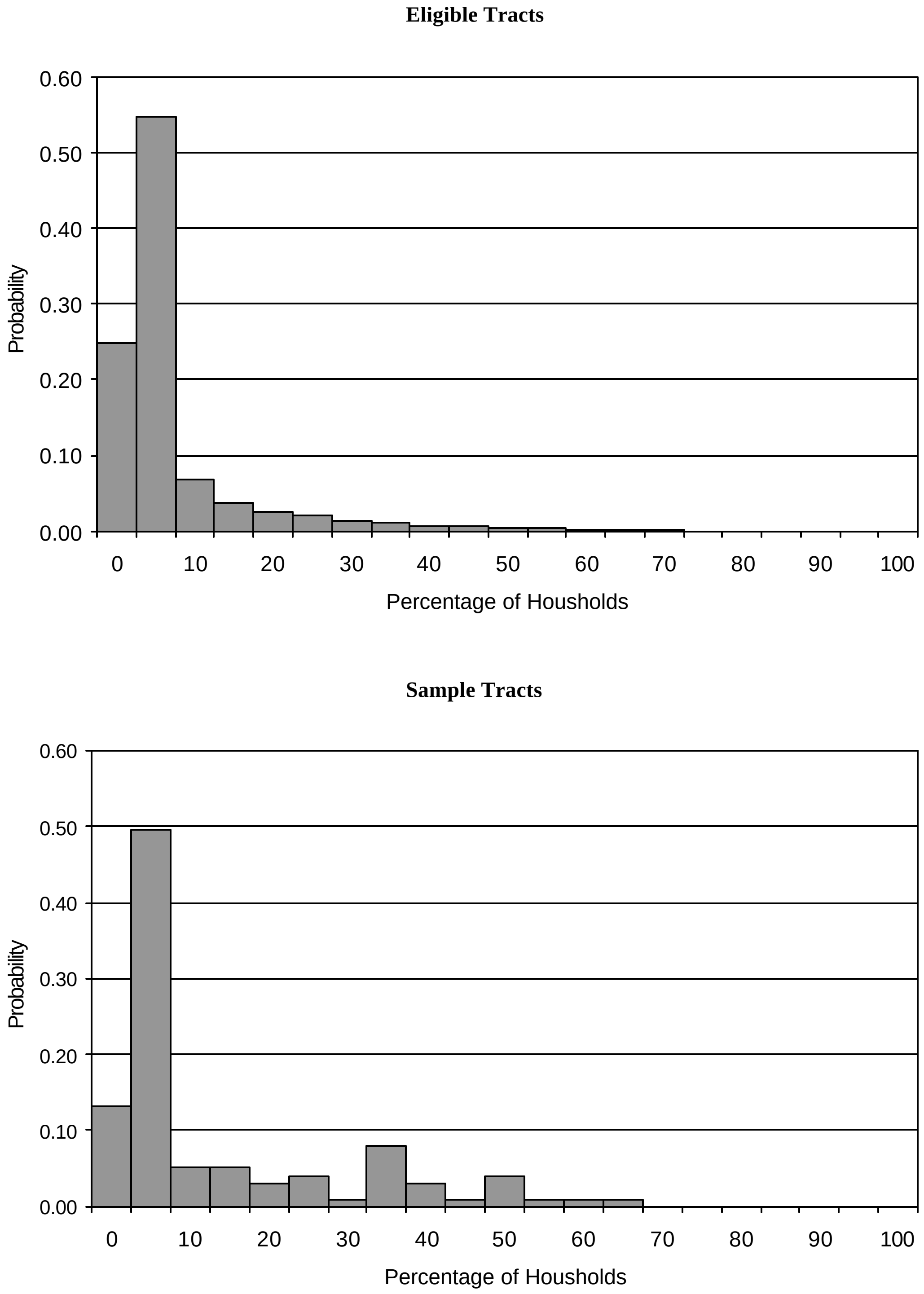


Figure 16. Percentage of Housing Units that Are Part-Time Residences

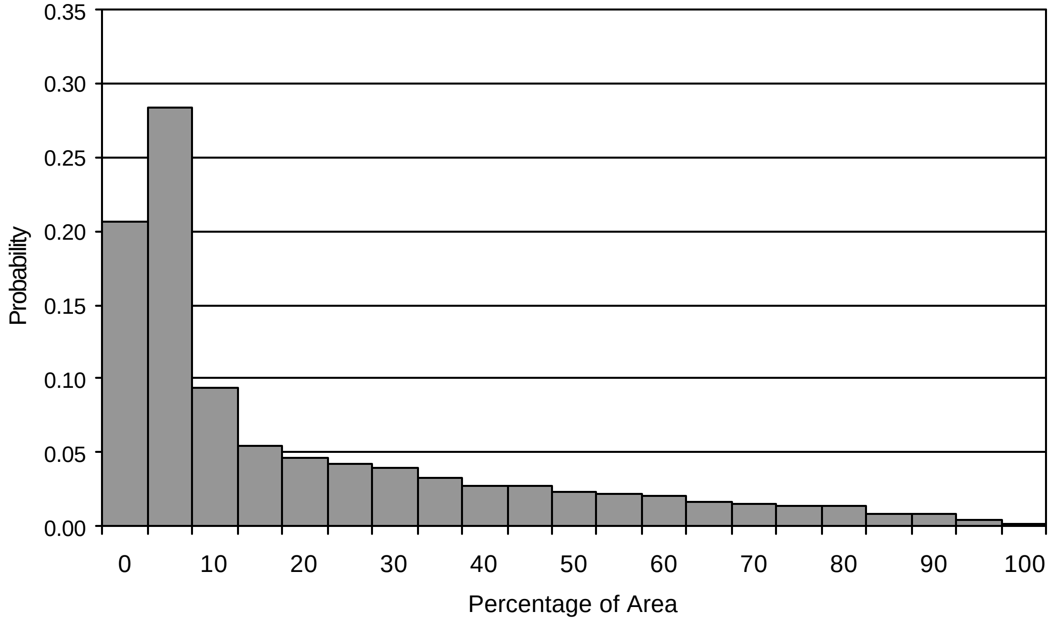


Eligible Tracts

Sample Tracts

Figure 17. Percentage of Area with Bedrock within 5 Feet of Surface

Eligible Tracts



Sample Tracts

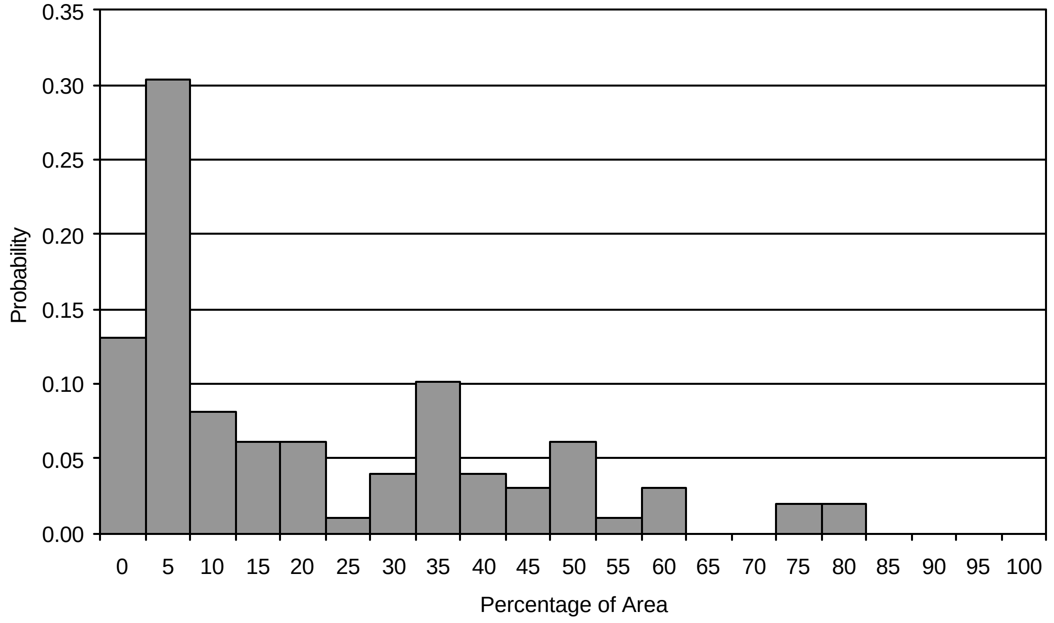
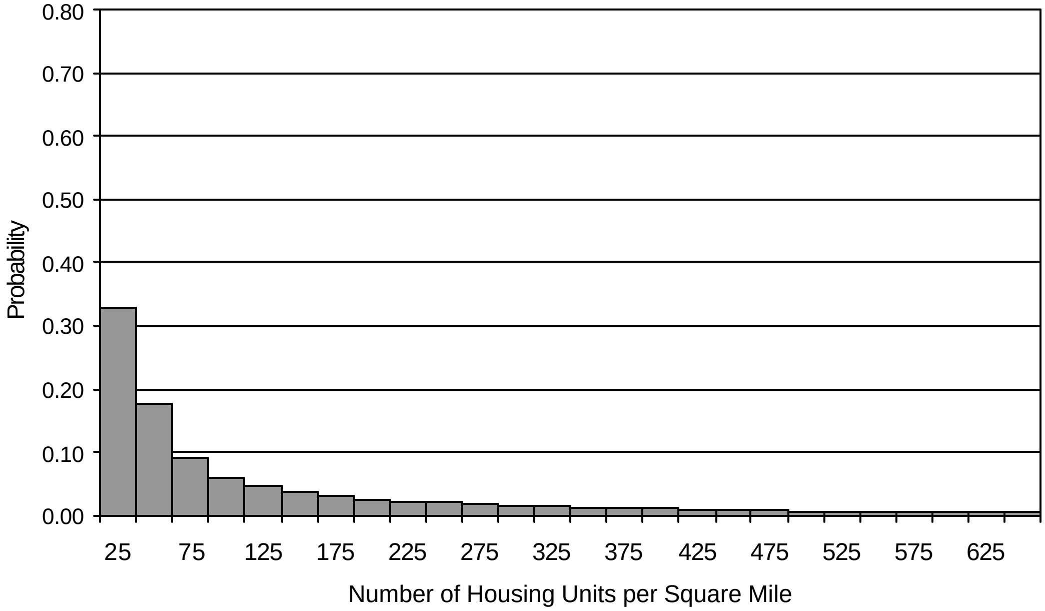


Figure 18. Total Housing Units per Square Mile

Eligible Tracts



Sample Tracts

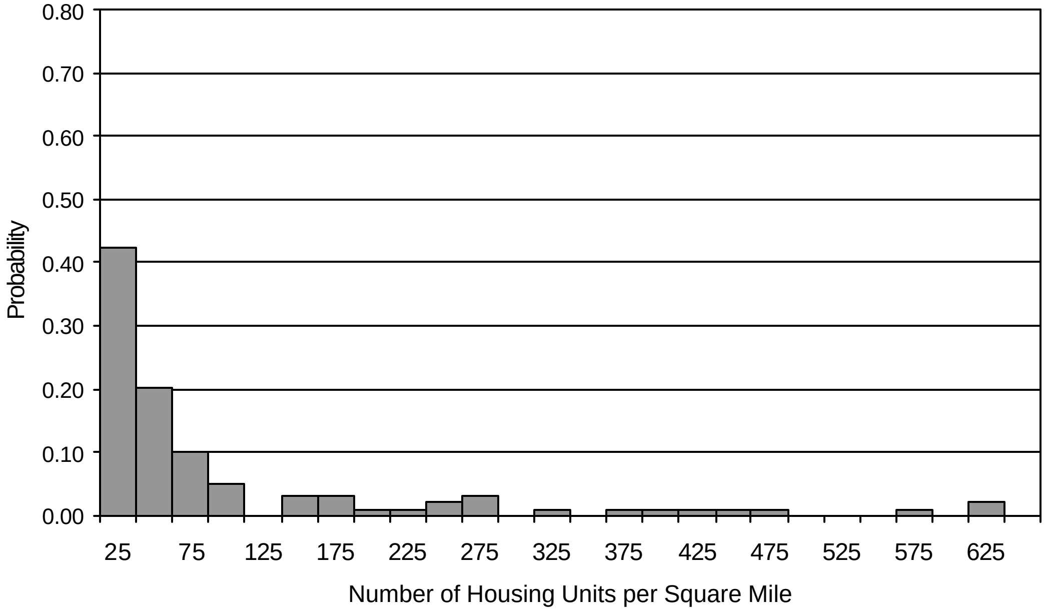
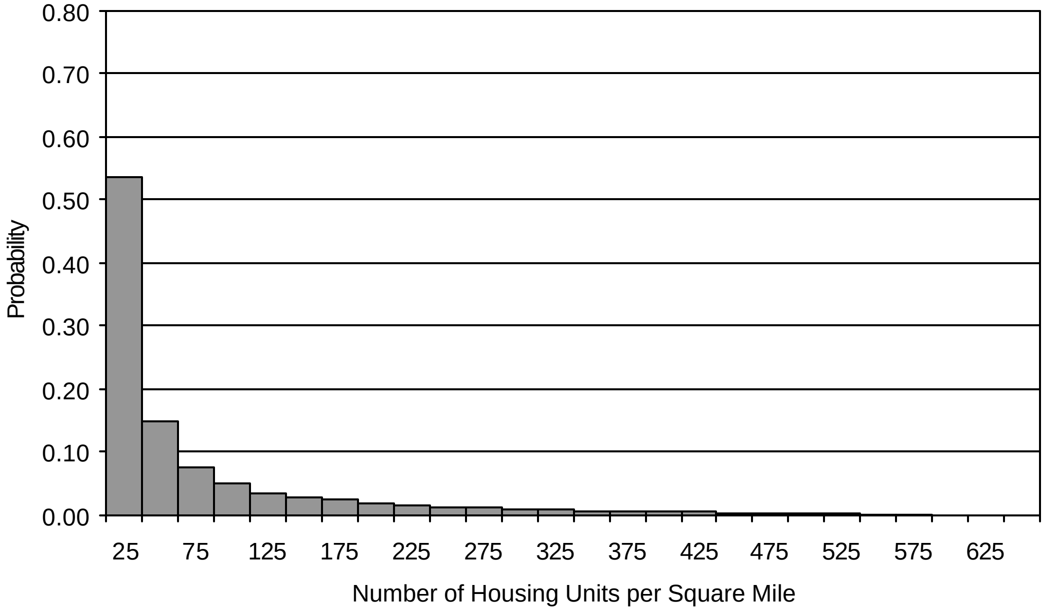


Figure 19. Housing Units Built before 1970 per Square Mile

Eligible Tracts



Sample Tracts

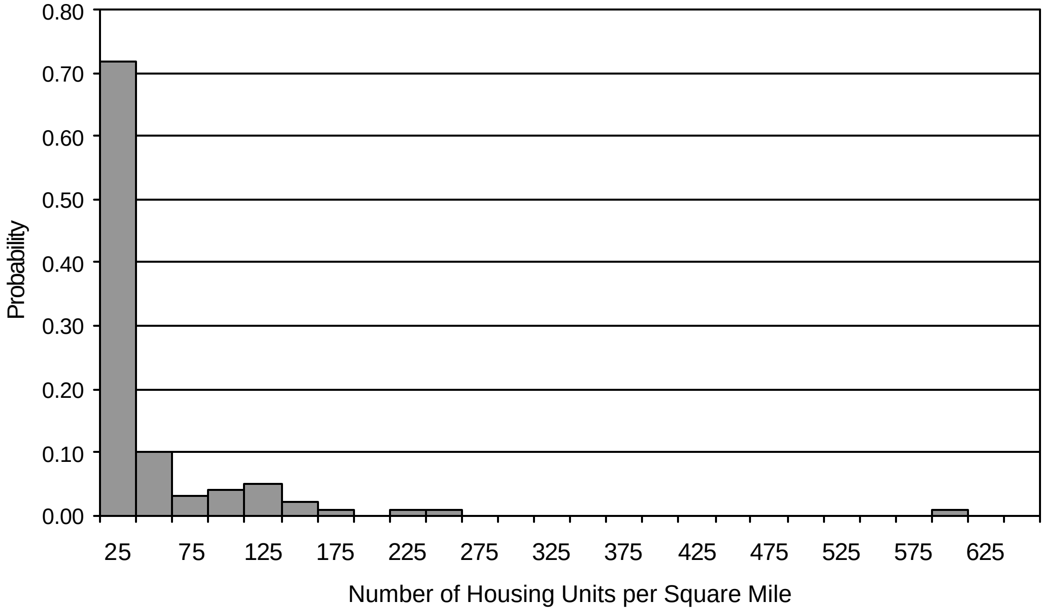
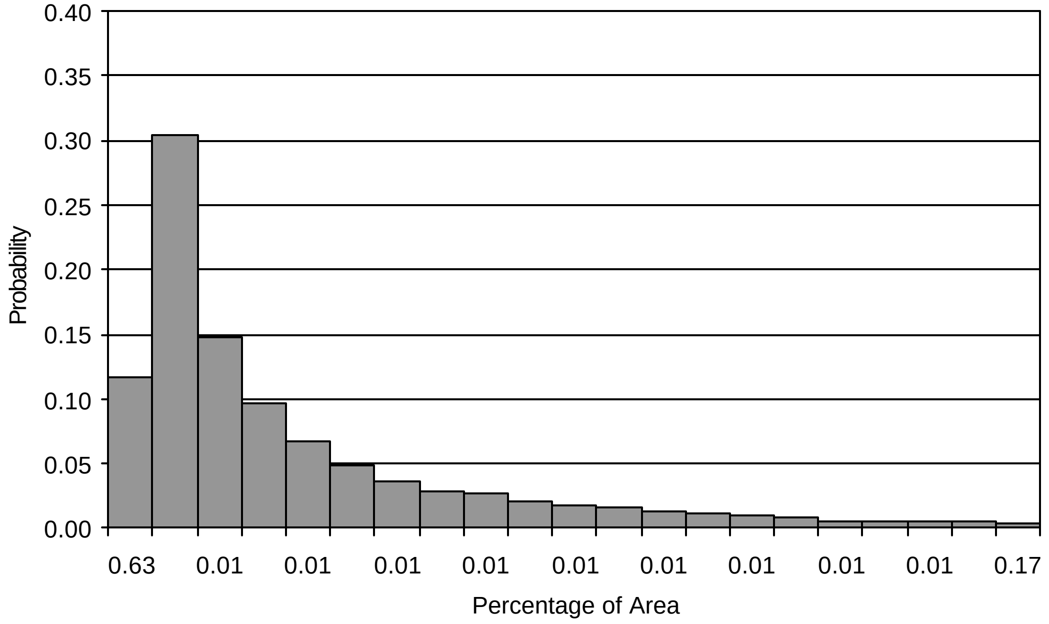


Figure 20. Percentage of Area with Poorly Drained Soil

Eligible Tracts



Sample Tracts

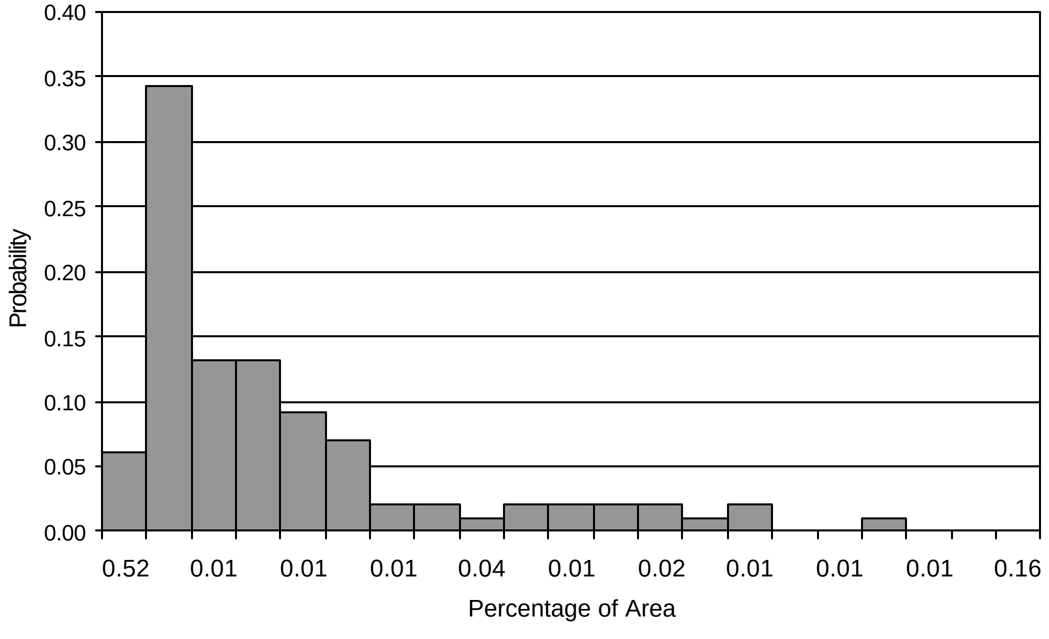
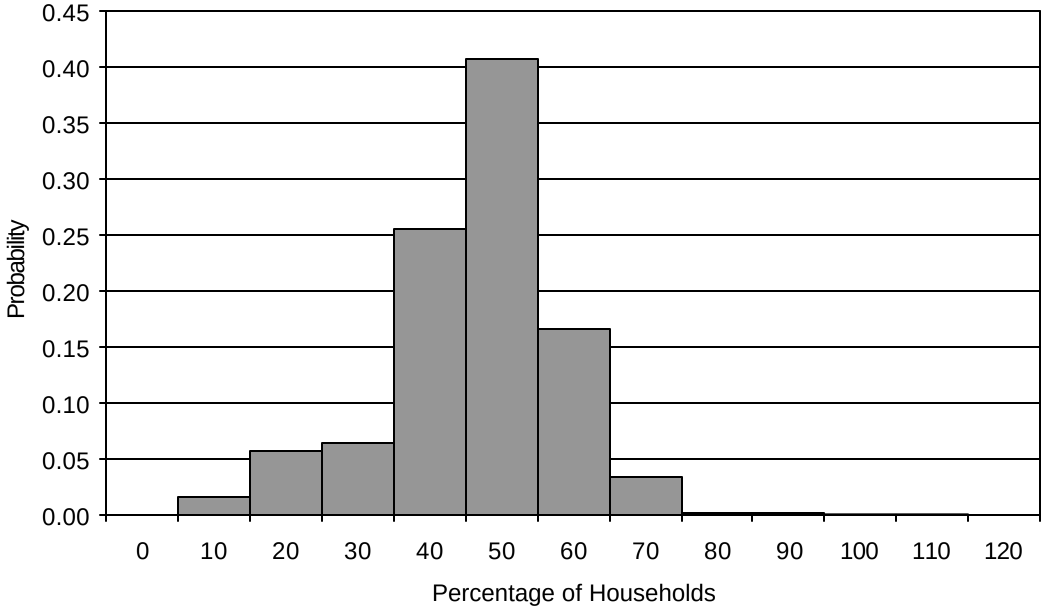
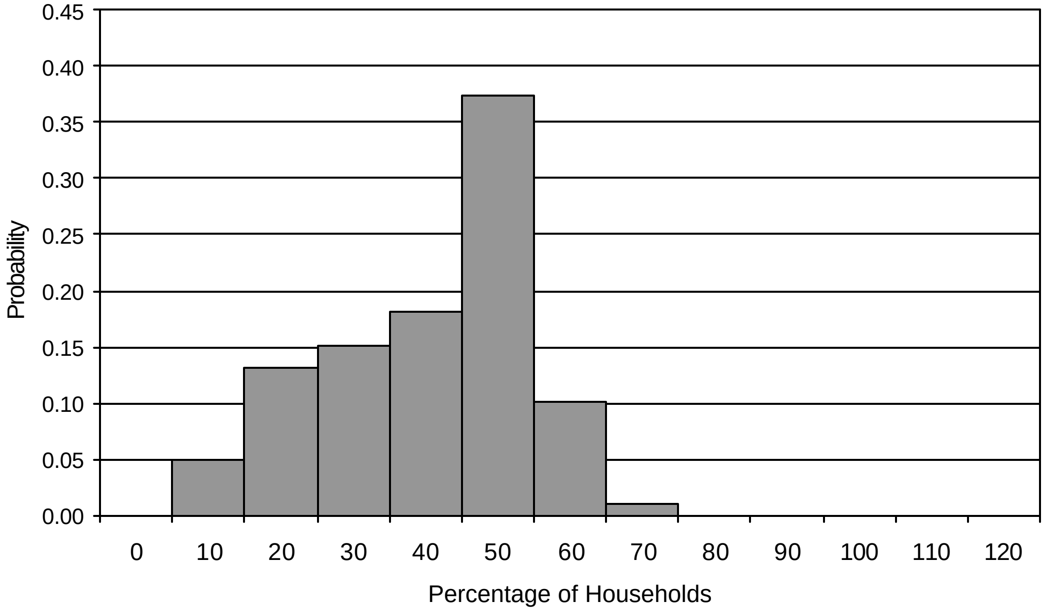


Figure 21. Mean Annual Precipitation

Eligible Tracts



Sample Tracts



REFERENCES

Duan, N., W.G. Manning, Jr., C.N. Morris, J.P. Newhouse. 1983. “A Comparison of Alternative Models for the Demand for Medical Care,” *Journal of Business and Economic Statistics,* Vol. 1., No. 2, April 1983.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap,* New York: Chapman and Hall, 1993.

Kott, P.S. 1994. "A Hypothesis Test of Linear Regression Coefficients with Survey Data," *Survey Methodology*, 20, 1994.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models,* New York: Chapman and Hall, 1989.

Snedecor, G.W. and W.G. Cochran. 1980. *Statistical Methods*, Seventh Edition. Ames, IA: Iowa State University Press, 1980.

U.S. EPA. 1994. *Guidance for the Data Quality Objectives Process*, EPA/600/R-96/055, EPA Quality Assurance Management Staff, Washington, 1994.